Anomalous datasets reveal metagenomic fabrication pipeline that further questions the legitimacy of RaTG13 genome and the associated Nature paper

John F. Signus

ABSTRACT

Recently, Daoyu Zhang et. Al [1], Mona Ralker et. Al [2] and Mohit Singla et. al [3] have reported exceptional anomalies associated with the RaTG13 metagenomic dataset which was inconsistent with that of a real fecal sample. Despite extensive search, we are only able to isolate 2 datasets Other than RaTG13 itself, that possessed significant levels of non-adaptor repeat sequences and absence of bacteria in the context of “bat” and “fecal” or “virome”.

Furthermore, the analysis of such datasets have revealed an established pipeline of which a viral sequence is “rehosted”—e.g. added to a metagenomic sample that originally did not contain such viral sequences. This raises serious questions to the legitimacy of RaTG13 genome and the associated Nature paper.

METHODS

Datasets

The NCBI SRA database was extensively searched using the term “bat” and “gut metagenome”, “feces”, “fecal” or “viral metagenome”. Datasets were first analyzed using NCBI TRACE, and the first 100 reads from the datasets were then analyzed for telomere-like repeats in the reads.

We only obtained 2 datasets with significant levels of telomere-like repeats and absence of bacteria.

Analysis using the SERRATUS toolbox

The 2 anomalous datasets obtained were subjected to multi-genome-wide alignment using the SERRATUS toolbox[4], which have been proven to be highly sensitive and is able to extract reads with potential alignments to all discovered or potential viral genomes known on NCBI, including very weak and partial alignments.

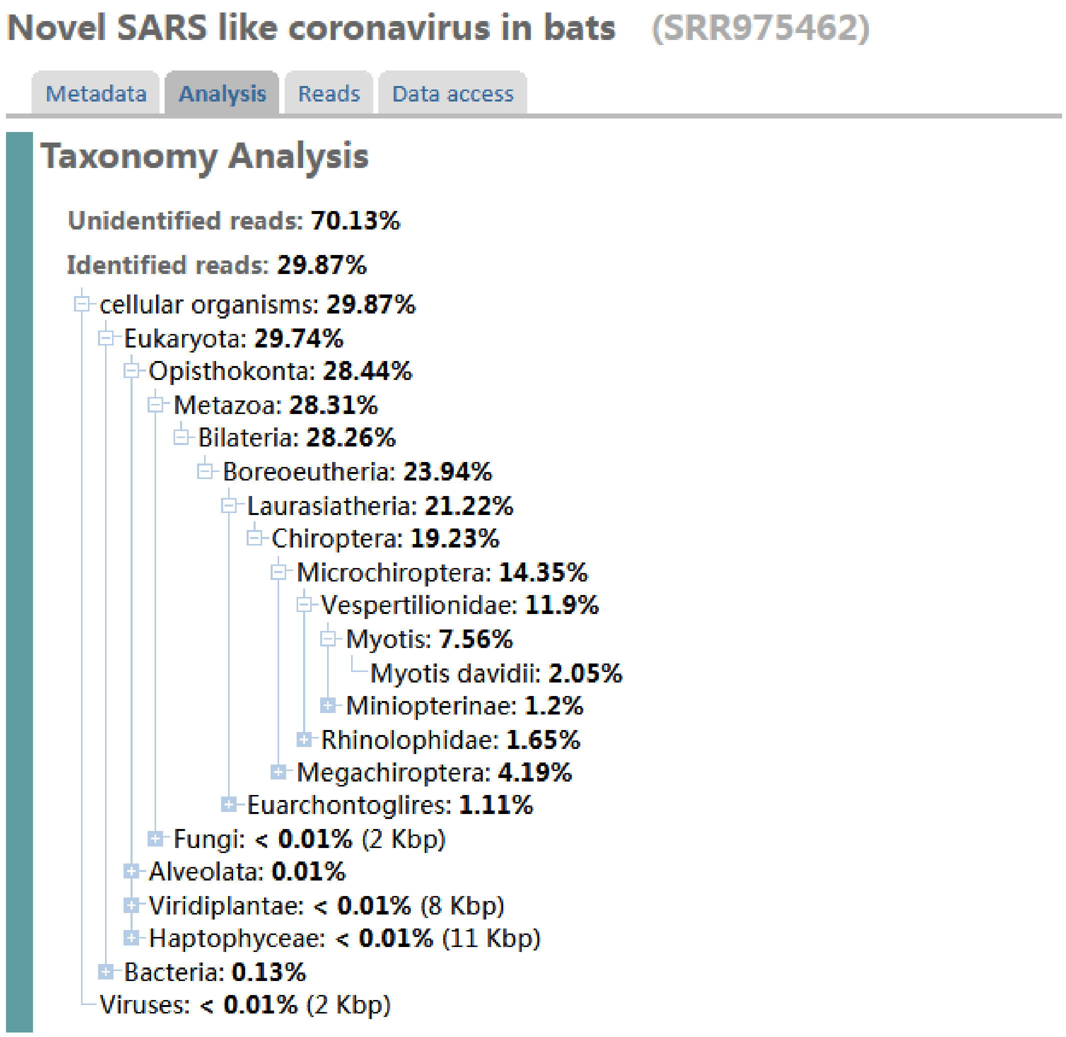
Reads extracted using the SERRATUS toolbox was then individually BLASTed on NCBI to exclude false positives, and the level of genome coverage was assayed for the likelihood of successful genome assembly.

RESULTS

Anomalous datasets obtained via extensive searching

Despite our effort of extensive searching on NCBI SRA, we obtained only 2 datasets that contained the reported anomalies by Zhang et al, Mona Ralker et. al and Mohit Singla et.al.
The datasets have accession number SRR975462 and SRR9644024, which contained 2% and 16% such sequences respectively.

Absence of assemblable viral RNA sequences in SRR975462



Novel SARS like coronavirus in bats (SRR975462)

Taxonomy Analysis

Unidentified reads: 70.13%

Identified reads: 29.87%

- cellular organisms: **29.87%**
  - Eukaryota: **29.74%**
    - Opisthokonta: **28.44%**
      - Metazoa: **28.31%**
        - Bilateria: **28.26%**
          - Boreoeutheria: **23.94%**
            - Laurasiatheria: **21.22%**
              - Chiroptera: **19.23%**
                - Microchiroptera: **14.35%**
                  - Vespertilionidae: **11.9%**
                    - Myotis: **7.56%**
                      - Myotis davidii: **2.05%**
                    - Miniopterinae: **1.2%**
                  - Rhinolophidae: **1.65%**
                - Megachiroptera: **4.19%**
            - Euarchontoglires: **1.11%**
      - Fungi: < **0.01%** (2 Kbp)
    - Alveolata: **0.01%**
    - Viridiplantae: < **0.01%** (8 Kbp)
    - Haptophyceae: < **0.01%** (11 Kbp)
  - Bacteria: **0.13%**
- Viruses: < **0.01%** (2 Kbp)

Fig.1: NCBI analysis of SRR975462.

Using the SERRATUS toolbox[8], a total of 19 reads from Coronaviridae covering 12% pangenome, 5 reads of Rhabdoviridae covering 4% pangenome, 1 single read from Astroviridae and 1 single read from picoRNAviridae was recovered.
None of these reads formed contiguous sequences with other reads, and no assembly either full or partial could be obtained from these sequences.

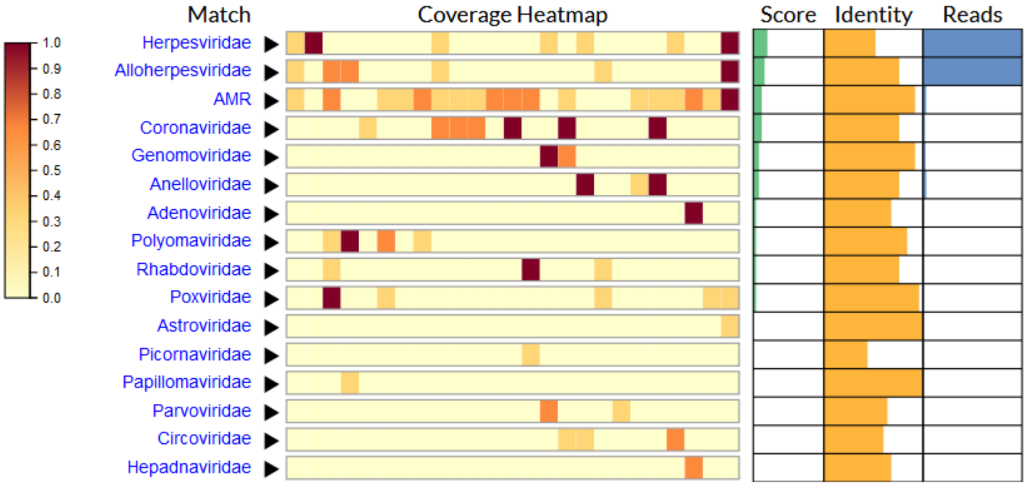


Fig.2: the SERRATUS result of SRR975462.

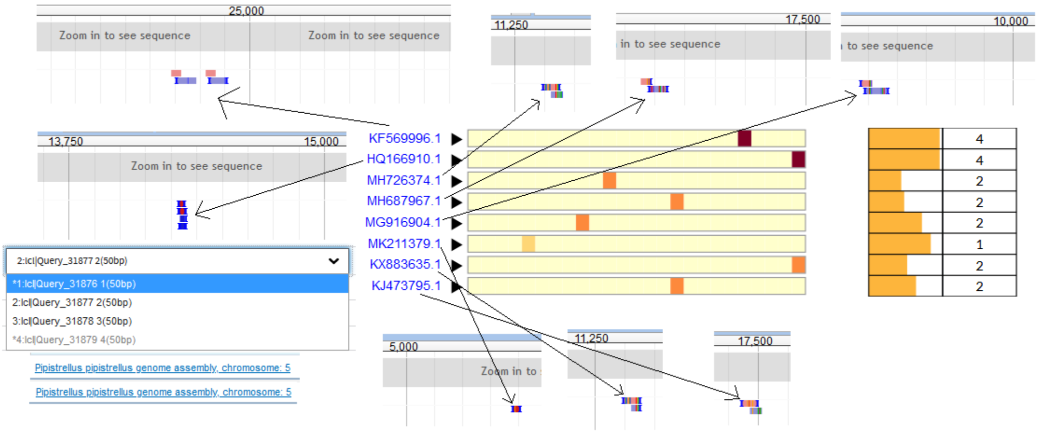


Fig.3a: The reads aligned to Coronaviridae in SRR975462. In addition to the fact that none of these reads formed a contiguous sequence, several of these reads were misaligned and when BLASTed as a whole, revealed to be nothing but a fragment of bat genomic DNA.

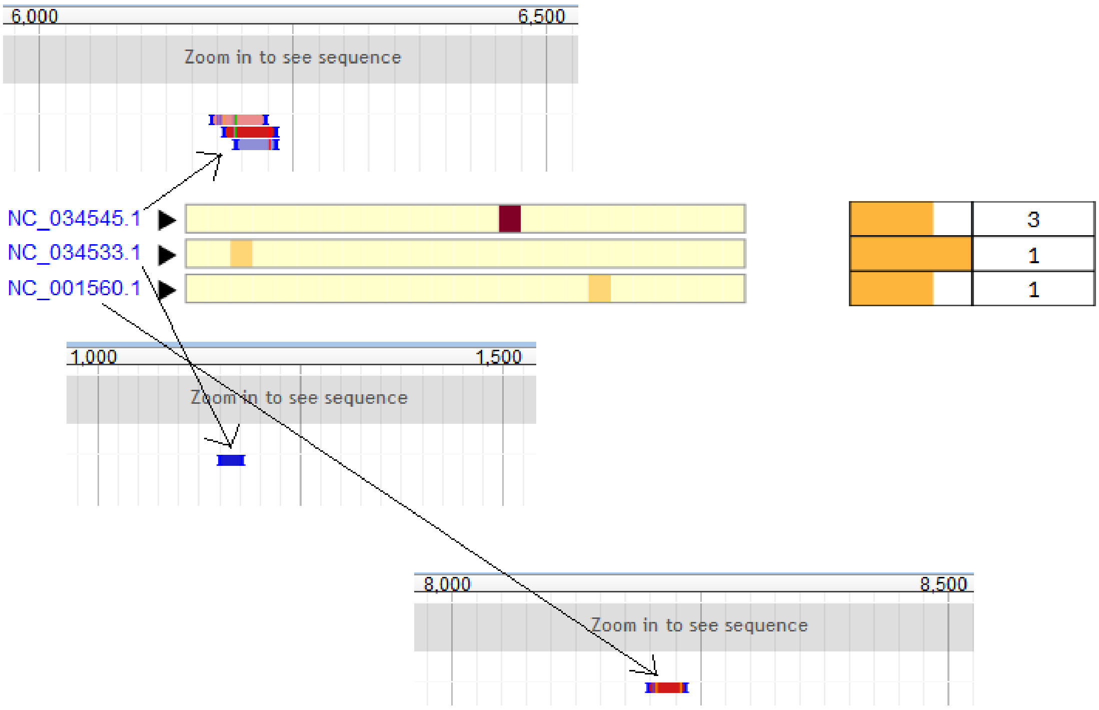
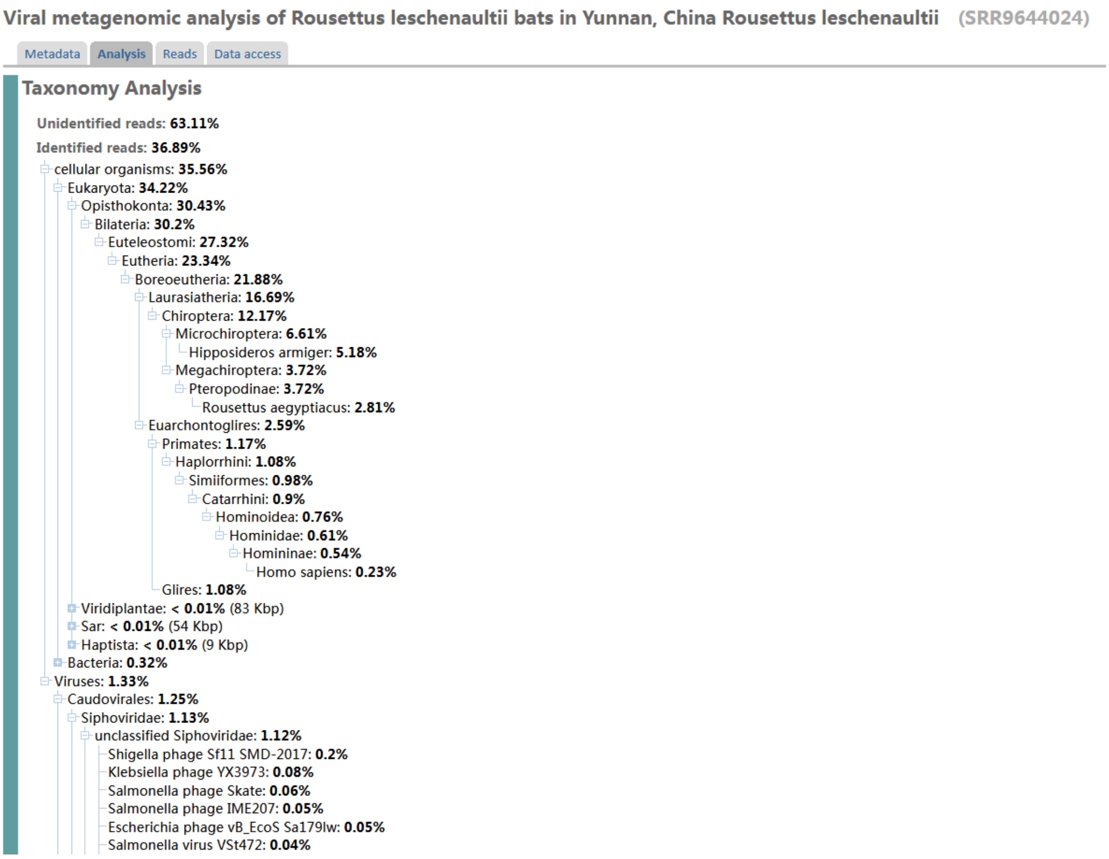


Fig.3b: The reads aligned to Rhabdoviridae in SRR975462. We did not obtain any meaningful assembly from such sequences.

SRR9644024 is a mixed dataset that does not match it’s description.



Viral metagenomic analysis of Rousettus leschenaultii bats in Yunnan, China Rousettus leschenaultii (SRR9644024)

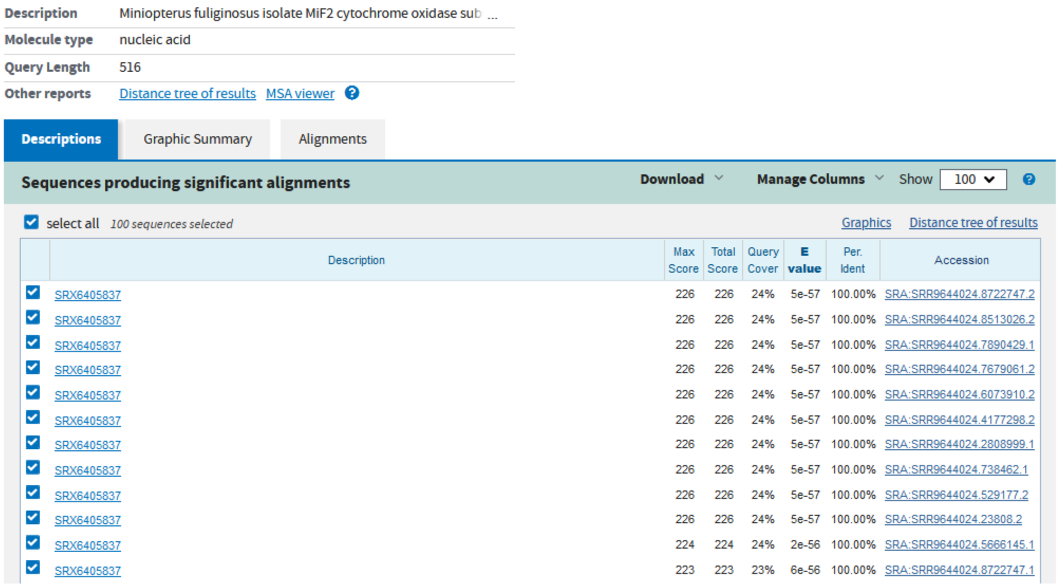
Taxonomy Analysis

```
Unidentified reads: 63.11%
Identified reads: 36.89%
cellular organisms: 35.56%
  Eukaryota: 34.22%
    Opisthokonta: 30.43%
      Bilateria: 30.2%
        Euteleostomi: 27.32%
          Eutheria: 23.34%
            Boreoeutheria: 21.88%
              Laurasiatheria: 16.69%
                Chiroptera: 12.17%
                  Microchiroptera: 6.61%
                    Hipposideros armiger: 5.18%
                  Megachiroptera: 3.72%
                    Pteropodinae: 3.72%
                      Rousettus aegyptiacus: 2.81%
              Euarchontoglires: 2.59%
                Primates: 1.17%
                  Haplorrhini: 1.08%
                    Simiiformes: 0.98%
                      Catarrhini: 0.9%
                        Hominoidea: 0.76%
                          Hominidae: 0.61%
                            Homininae: 0.54%
                              Homo sapiens: 0.23%
                Glires: 1.08%
    Viridiplantae: < 0.01% (83 Kbp)
    Sar: < 0.01% (54 Kbp)
    Haptista: < 0.01% (9 Kbp)
  Bacteria: 0.32%
Viruses: 1.33%
  Caudovirales: 1.25%
    Siphoviridae: 1.13%
      unclassified Siphoviridae: 1.12%
        Shigella phage Sf11 SMD-2017: 0.2%
        Klebsiella phage YX3973: 0.08%
        Salmonella phage Skate: 0.06%
        Salmonella phage IME207: 0.05%
        Escherichia phage vB_EcoS Sa179lw: 0.05%
        Salmonella virus VSt472: 0.04%
```

Fig.4: The NCBI analysis results of SRR9644024 and the description of the dataset.

Despite the dataset and the associated BioSample claim a host of Rousettus Leschunatii, the NCBI analysis revealed multiple bat species including Hipposideros Armiger and Miniopterus Natalensis, as well as a sizable fraction of Homo Sapiens.

By using Mitochondrial genome and COI genes, we are able to isolate from the dataset material from Hipposideros Armiger, Miniopterus fuliginosus, Rhinolophus Affinis, Rhinolophus Pearsonii, Rhinolophus Monoceros and Homo Sapiens.



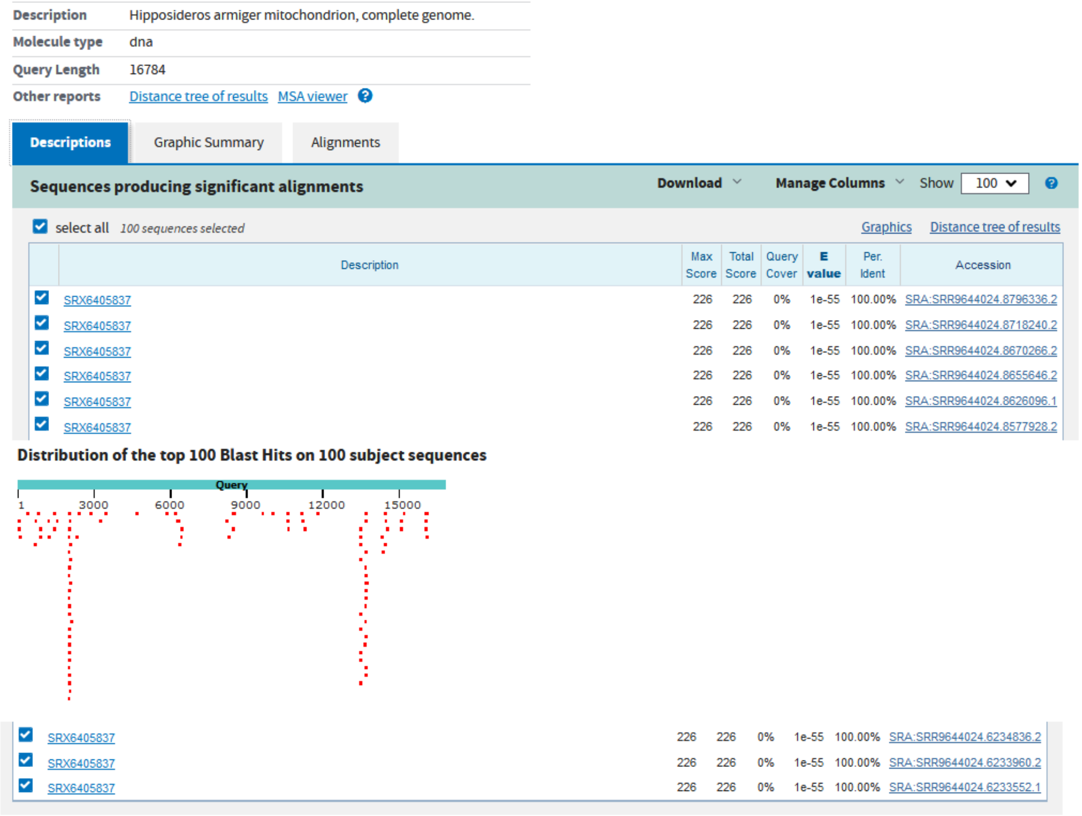
Description	Miniopterus fuliginosus isolate MiF2 cytochrome oxidase sub ...
Molecule type	nucleic acid
Query Length	516
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.8722747.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.8513026.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.7890429.1
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.7679061.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.6073910.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.4177298.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.2808999.1
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.738462.1
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.529177.2
☑	SRX6405837	226	226	24%	5e-57	100.00%	SRA:SRR9644024.23808.2
☑	SRX6405837	224	224	24%	2e-56	100.00%	SRA:SRR9644024.5666145.1
☑	SRX6405837	223	223	23%	6e-56	100.00%	SRA:SRR9644024.8722747.1

Fig.5a: Miniopterus fuliginosus Cytochrome Oxidase 1 reads recovered from SRR9644024



Description	Hipposideros armiger mitochondrion, complete genome.
Molecule type	dna
Query Length	16784
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

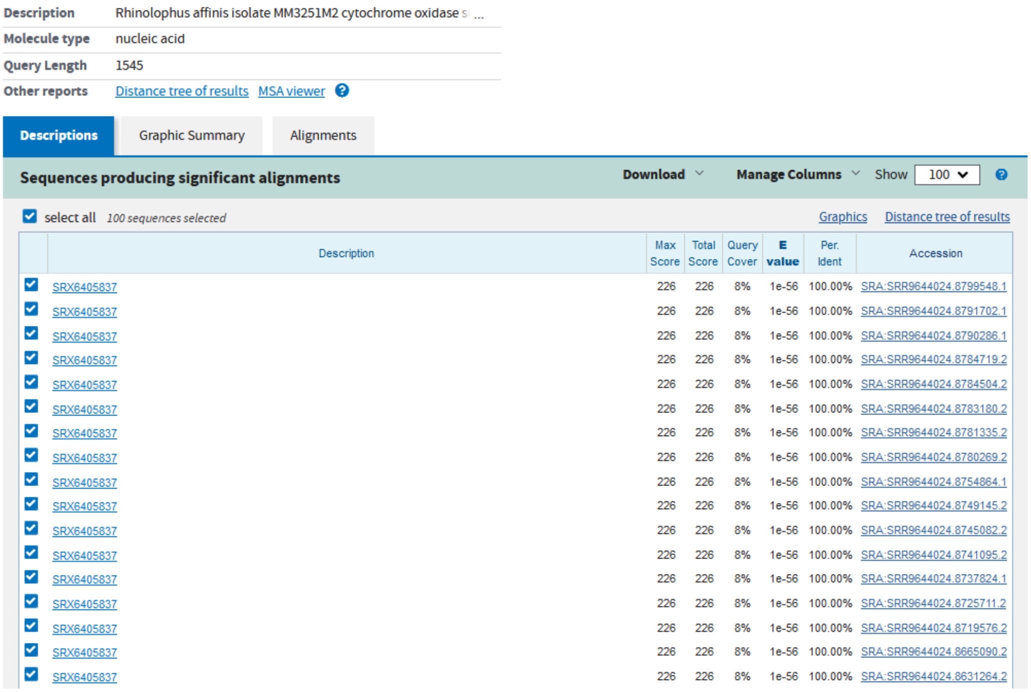
☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8796336.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8718240.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8670266.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8655646.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8626096.1
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.8577928.2

Distribution of the top 100 Blast Hits on 100 subject sequences

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.6234836.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.6233960.2
☑	SRX6405837	226	226	0%	1e-55	100.00%	SRA:SRR9644024.6233552.1

Fig.5b: Hipposideros armiger 100% full-length matched mitogenome recovered from SRR9644024



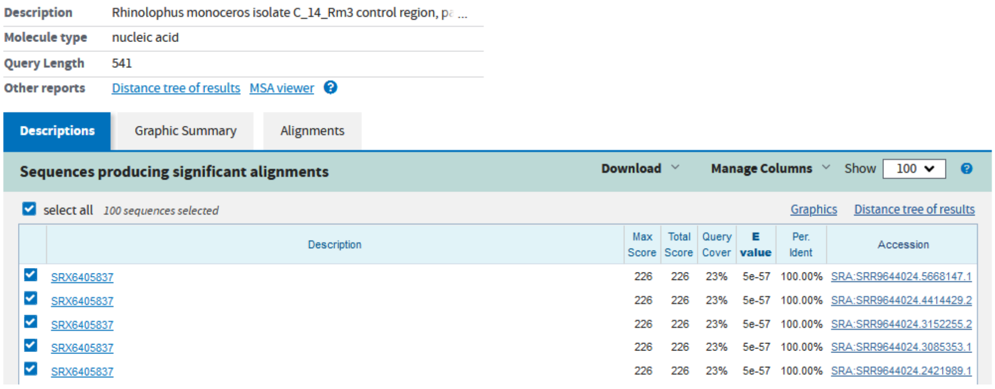
Description	Rhinolophus affinis isolate MM3251M2 cytochrome oxidase s ...
Molecule type	nucleic acid
Query Length	1545
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8799548.1
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8791702.1
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8790286.1
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8784719.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8784504.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8783180.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8781335.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8780269.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8754864.1
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8749145.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8745082.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8741095.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8737824.1
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8725711.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8719576.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8665090.2
☑	SRX6405837	226	226	8%	1e-56	100.00%	SRA:SRR9644024.8631264.2

Fig.5c: Rhinolophus affinis Cytochrome Oxidase I reads recovered from SRR9644024



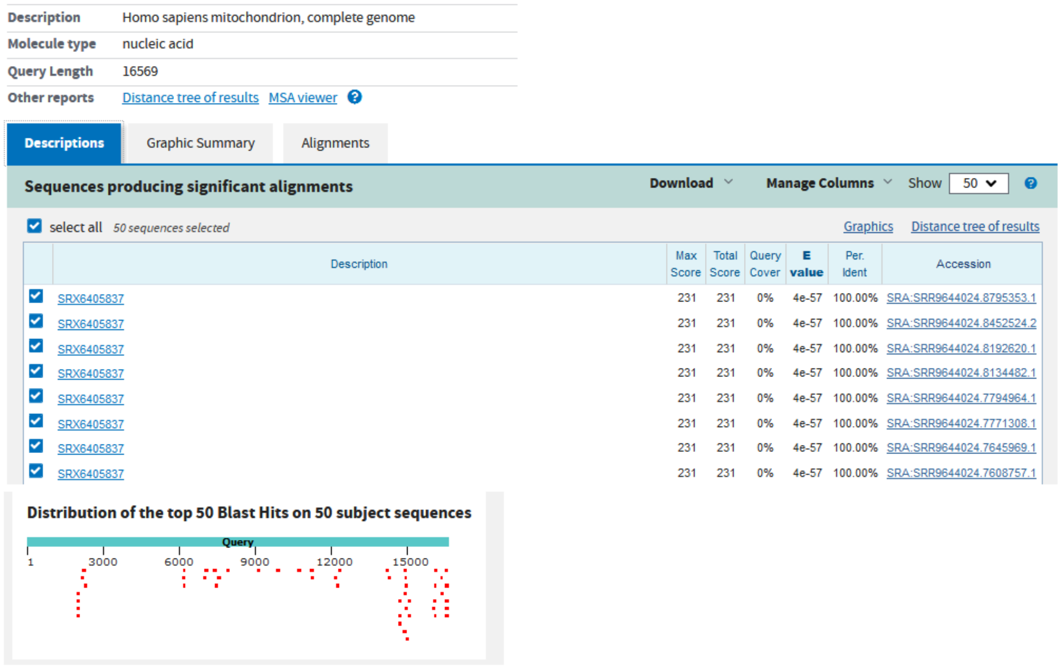
Description	Rhinolophus monoceros isolate C_14_Rm3 control region, pa ...
Molecule type	nucleic acid
Query Length	541
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	226	226	23%	5e-57	100.00%	SRA:SRR9644024.5668147.1
☑	SRX6405837	226	226	23%	5e-57	100.00%	SRA:SRR9644024.4414429.2
☑	SRX6405837	226	226	23%	5e-57	100.00%	SRA:SRR9644024.3152255.2
☑	SRX6405837	226	226	23%	5e-57	100.00%	SRA:SRR9644024.3085353.1
☑	SRX6405837	226	226	23%	5e-57	100.00%	SRA:SRR9644024.2421989.1

Fig.5d: Rhinolophus Monoceros Mitochondrial D-loop reads recovered from SRR9644024



Description	Homo sapiens mitochondrion, complete genome
Molecule type	nucleic acid
Query Length	16569
Other reports	Distance tree of results MSA viewer ?

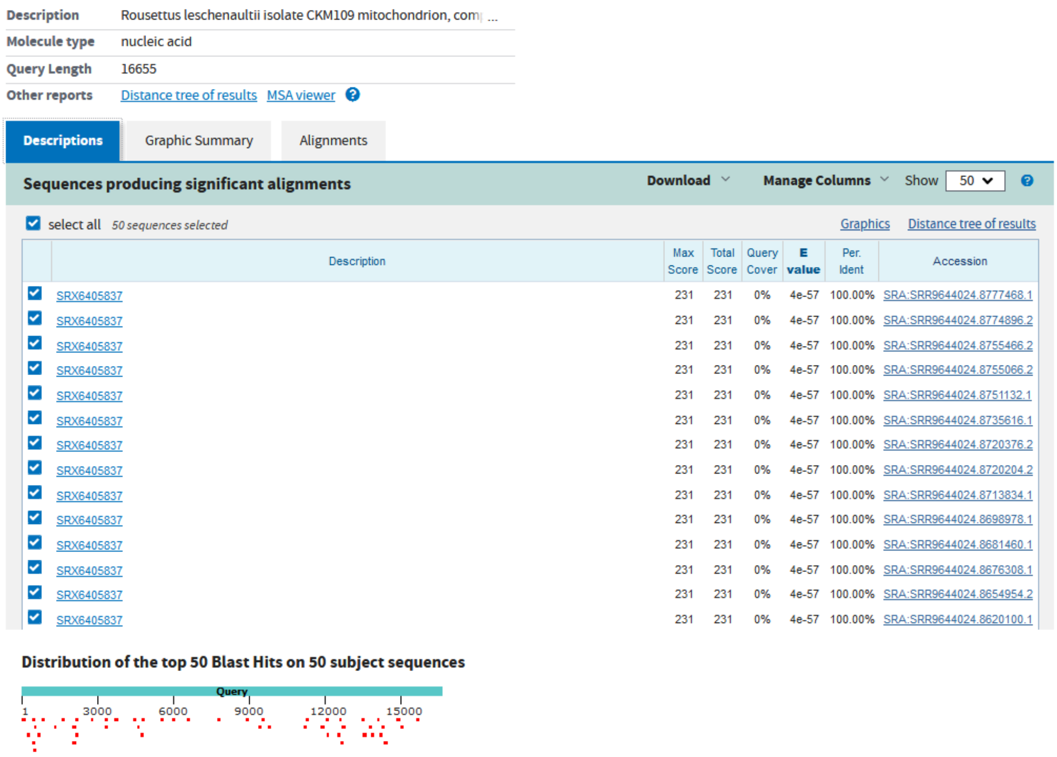
Sequences producing significant alignments

☑ select all *50 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8795353.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8452524.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8192620.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8134482.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.7794964.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.7771308.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.7645969.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.7608757.1

Distribution of the top 50 Blast Hits on 50 subject sequences

Fig.5e:Homo Sapiens 100% full-length matched Mitogenome recovered from SRR9644024



Description	Rousettus leschenaultii isolate CKM109 mitochondrion, com| ...
Molecule type	nucleic acid
Query Length	16655
Other reports	Distance tree of results MSA viewer ?

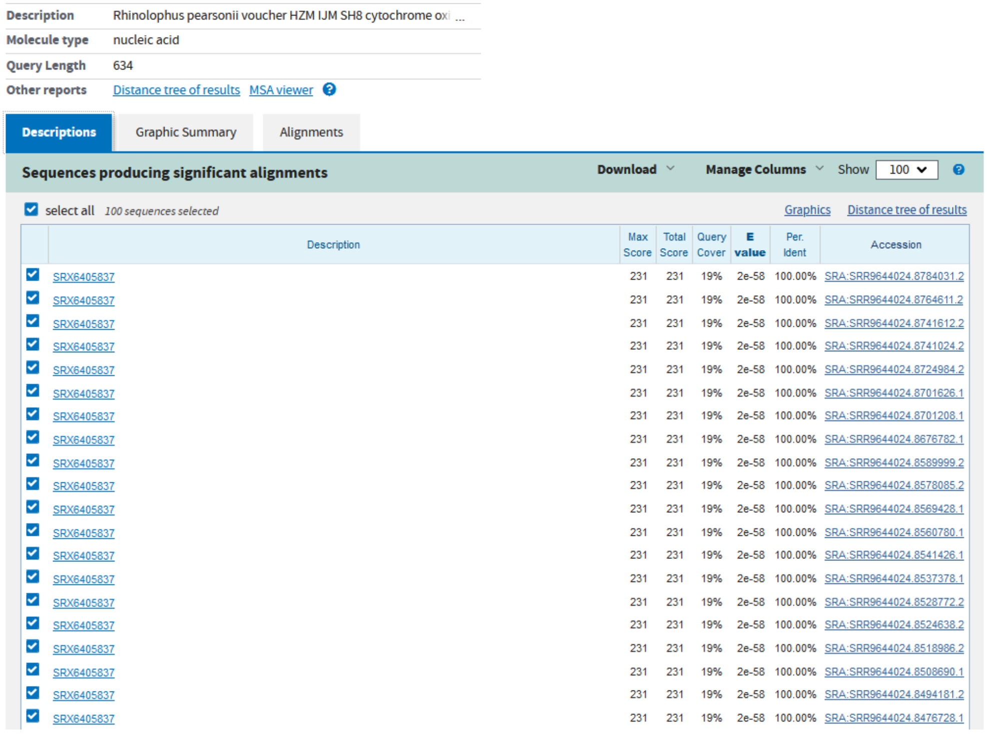
Sequences producing significant alignments

☑ select all *50 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8777468.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8774896.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8755466.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8755066.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8751132.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8735616.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8720376.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8720204.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8713834.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8698978.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8681460.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8676308.1
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8654954.2
☑	SRX6405837	231	231	0%	4e-57	100.00%	SRA:SRR9644024.8620100.1

Distribution of the top 50 Blast Hits on 50 subject sequences

Fig.5f: Rousettus leschenaultii Mitogenome recovered from SRR9644024



Description	Rhinolophus pearsonii voucher HZM IJM SH8 cytochrome ox ...
Molecule type	nucleic acid
Query Length	634
Other reports	Distance tree of results MSA viewer

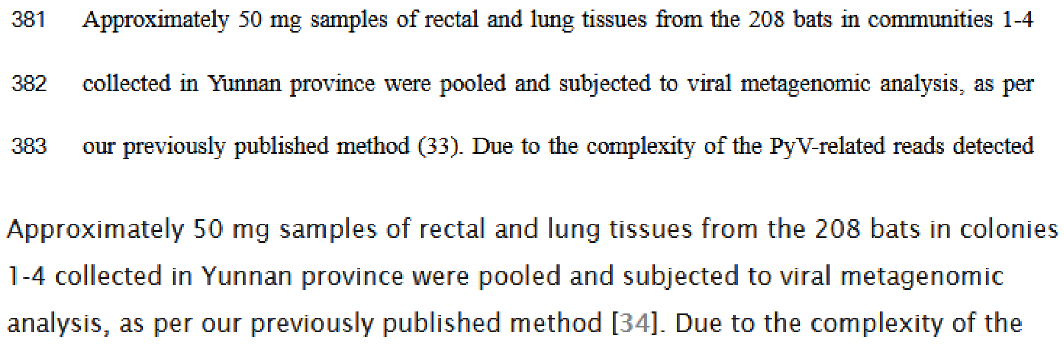
Sequences producing significant alignments

☑ select all 100 sequences selected

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8784031.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8764611.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8741612.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8741024.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8724984.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8701626.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8701208.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8676782.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8589999.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8578085.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8569428.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8560780.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8541426.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8537378.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8528772.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8524638.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8518986.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8508690.1
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8494181.2
☑	SRX6405837	231	231	19%	2e-58	100.00%	SRA:SRR9644024.8476728.1

Fig.5g: Rhinolophus Pearsonii Cytochrome Oxidase I (COX1) 100% fully matched reads recovered from SRR9644024.

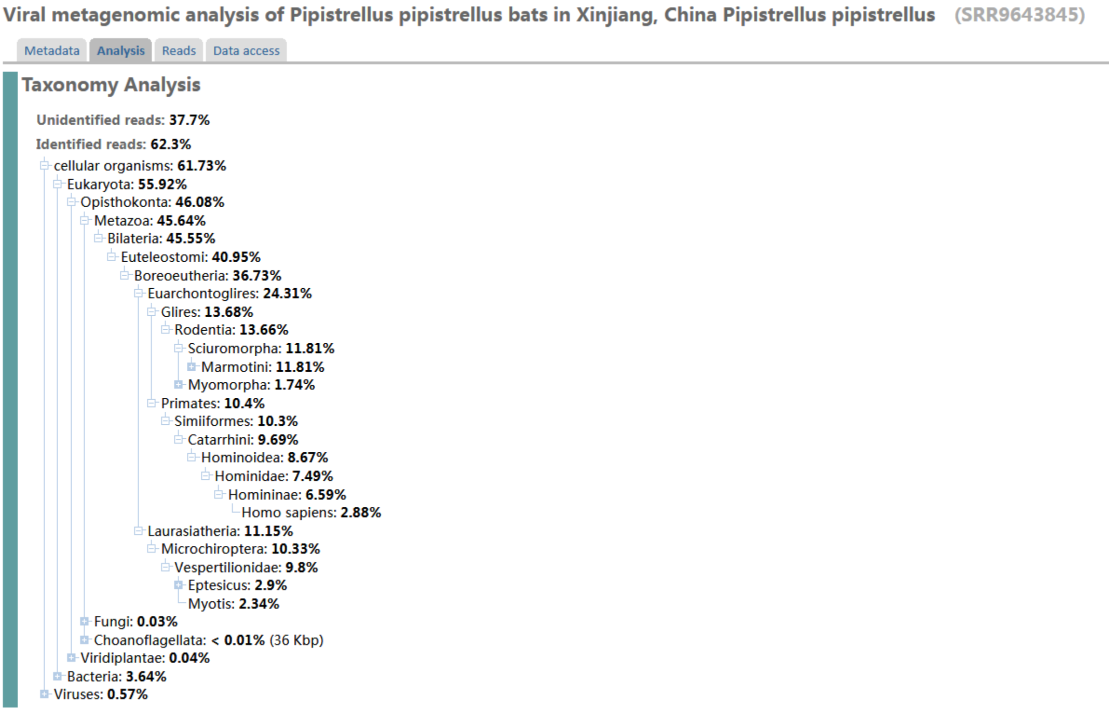
As SRR9644024 was supposed to be a sample from Rousettus Leschunatii, the presence of reads from a wide range of different bat species including both Microchiroptera and Megachiroptera was impossible even given exceptionally contaminated sample collection environment. In deed, the associated BiorXiV preprint[4] and JVM article[5] defines the sample as "pooled lung and rectal tissues" rather than "feces". Notably, the samples were "archived and sub-packed samples" which gives rise to the chance for accidental inclusion of experimental fabrication products and PCR products, as the related SRA dataset, SRR9643845, does not show evidence of any anomalies within the reads.



Approximately 50 mg samples of rectal and lung tissues from the 208 bats in communities 1-4 collected in Yunnan province were pooled and subjected to viral metagenomic analysis, as per our previously published method (33). Due to the complexity of the PyV-related reads detected

Approximately 50 mg samples of rectal and lung tissues from the 208 bats in colonies 1-4 collected in Yunnan province were pooled and subjected to viral metagenomic analysis, as per our previously published method [34]. Due to the complexity of the

Fig.6: the methods section from [4] and [5] showing the designation of the samples used in SRR9644024 as being tissue samples rather than feces.



Viral metagenomic analysis of Pipistrellus pipistrellus bats in Xinjiang, China Pipistrellus pipistrellus (SRR9643845)

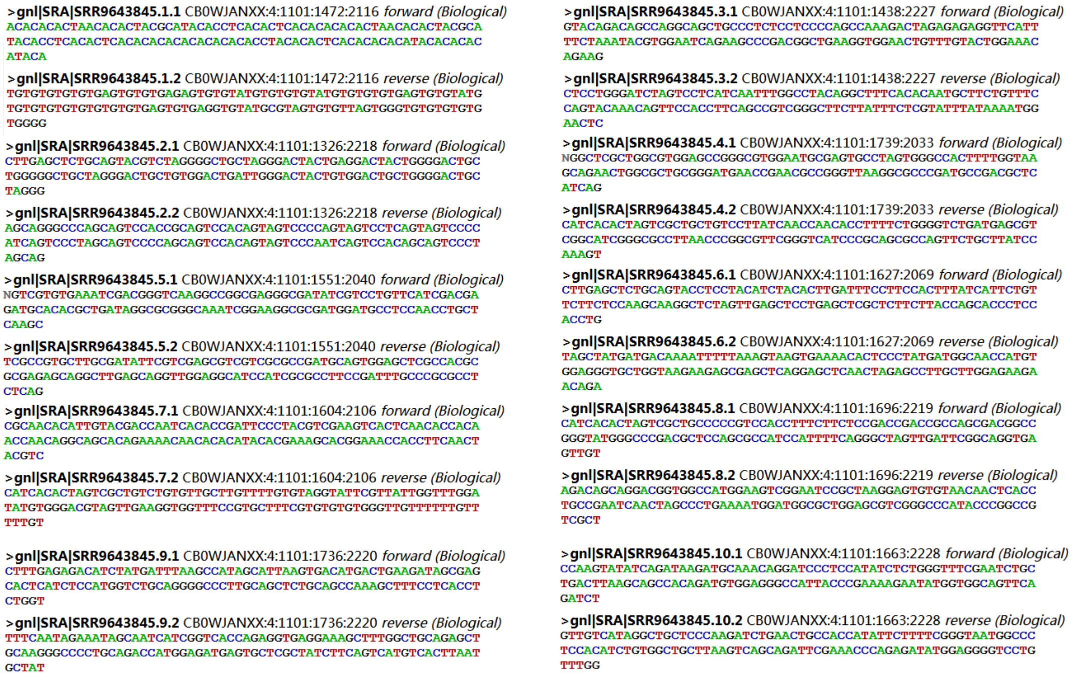
Taxonomy Analysis

Unidentified reads: **37.7%**

Identified reads: **62.3%**

```
cellular organisms: 61.73%
  Eukaryota: 55.92%
    Opisthokonta: 46.08%
      Metazoa: 45.64%
        Bilateria: 45.55%
          Euteleostomi: 40.95%
            Boreoeutheria: 36.73%
              Euarchontoglires: 24.31%
                Glires: 13.68%
                  Rodentia: 13.66%
                    Sciuromorpha: 11.81%
                      Marmotini: 11.81%
                    Myomorpha: 1.74%
                Primates: 10.4%
                  Simiiformes: 10.3%
                    Catarrhini: 9.69%
                      Hominoidea: 8.67%
                        Hominidae: 7.49%
                          Homininae: 6.59%
                            Homo sapiens: 2.88%
              Laurasiatheria: 11.15%
                Microchiroptera: 10.33%
                  Vespertilionidae: 9.8%
                    Eptesicus: 2.9%
                    Myotis: 2.34%
      Fungi: 0.03%
      Choanoflagellata: < 0.01% (36 Kbp)
    Viridiplantae: 0.04%
  Bacteria: 3.64%
Viruses: 0.57%
```

Fig.7: Analysis of SRR9643845. The Bacterial read percentage in total cellular organisms is 5.98%.

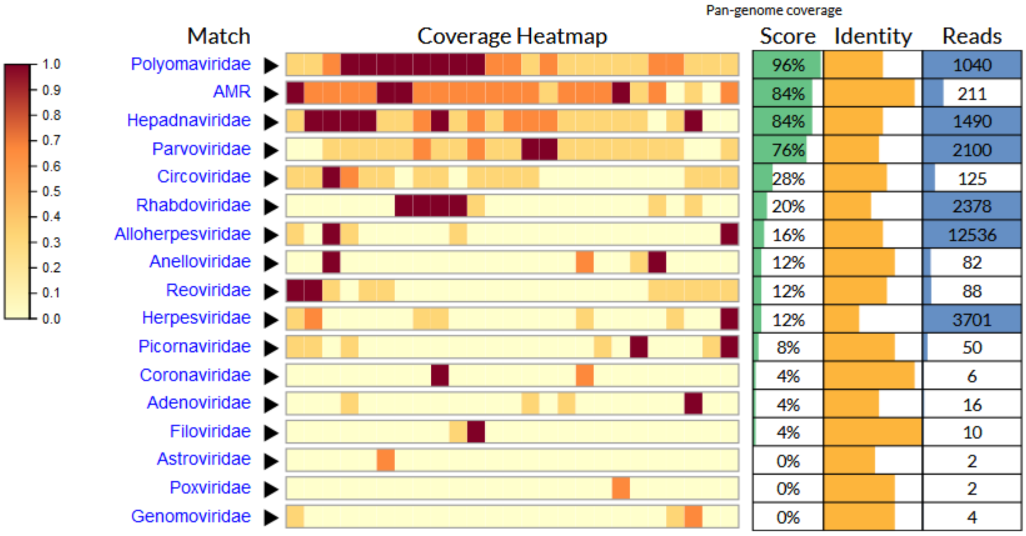


```
>gnl|SRA|SRR9643845.1.1 CB0WJANXX:4:1101:1472:2116 forward (Biological)
ACACACACTAACACACTACGCATACACCTCACACTCACACACACACTAACACACTACGCA
TACACCTCACACTCACACACACACACACACCTACACACTCACACACACATACACACAC
ATACA

>gnl|SRA|SRR9643845.1.2 CB0WJANXX:4:1101:1472:2116 reverse (Biological)
TGTGTGTGTGTGAGTGTGTGAGAGTGTGTATGTGTGTGTATGTGTGTGTGAGTGTGTATG
TGTGTGTGTGTGTGTGTGAGTGTGAGGTGTATGCGTAGTGTGTTAGTGGGTGTGTGTGTG
TGGGG

>gnl|SRA|SRR9643845.2.1 CB0WJANXX:4:1101:1326:2218 forward (Biological)
CTTGAGCTCTGCAGTACGTCTAGGGGCTGCTAGGGACTACTGAGGACTACTGGGGACTGC
TGGGGGCTGCTAGGGACTGCTGTGGACTGATTGGGACTACTGTGGACTGCTGGGGACTGC
TAGGG

>gnl|SRA|SRR9643845.2.2 CB0WJANXX:4:1101:1326:2218 reverse (Biological)
AGCAGGGCCCAGCAGTCCACCGCAGTCCACAGTAGTCCCCAGTAGTCCTCAGTAGTCCCC
ATCAGTCCCTAGCAGTCCCCAGCAGTCCACAGTAGTCCCAATCAGTCCACAGCAGTCCCT
AGCAG

>gnl|SRA|SRR9643845.5.1 CB0WJANXX:4:1101:1551:2040 forward (Biological)
NGTCGTGTGAAATCGACGGGTCAAGGCCGGCGAGGGCGATATCGTCCTGTTCATCGACGA
GATGCACACGCTGATAGGCGCGGGCAAATCGGAAGGCGCGATGGATGCCTCCAACCTGCT
CAAGC

>gnl|SRA|SRR9643845.5.2 CB0WJANXX:4:1101:1551:2040 reverse (Biological)
TCGCCGTGCTTGCGATATTCGTCGAGCGTCGTCGCGCCGATGCAGTGGAGCTCGCCACGC
GCGAGAGCAGGCTTGAGCAGGTTGGAGGCATCCATCGCGCCTTCCGATTTGCCCGCGCCT
CTCAG

>gnl|SRA|SRR9643845.7.1 CB0WJANXX:4:1101:1604:2106 forward (Biological)
CGCAACACATTGTACGACCAATCACACCGATTCCCTACGTCGAAGTCACTCAACACCACA
ACCAACAGGCAGCACAGAAAACAACACACATACACGAAAGCACGGAAACCACCTTCAACT
ACGTC

>gnl|SRA|SRR9643845.7.2 CB0WJANXX:4:1101:1604:2106 reverse (Biological)
CATCACACTAGTCGCTGTCTGTGTTGCTTGTTTTGTGTAGGTATTCGTTATTGGTTTGGA
TATGTGGGACGTAGTTGAAGGTGGTTTCCGTGCTTTCGTGTGTGTGGGTTGTTTTTTGTT
TTTGT

>gnl|SRA|SRR9643845.9.1 CB0WJANXX:4:1101:1736:2220 forward (Biological)
CTTTGAGAGACATCTATGATTTAAGCCATAGCATTAAGTGACATGACTGAAGATAGCGAG
CACTCATCTCCATGGTCTGCAGGGGCCCTTGCAGCTCTGCAGCCAAAGCTTTCCTCACCT
CTGGT

>gnl|SRA|SRR9643845.9.2 CB0WJANXX:4:1101:1736:2220 reverse (Biological)
TTTCAATAGAAATAGCAATCATCGGTCACCAGAGGTGAGGAAAGCTTTGGCTGCAGAGCT
GCAAGGGCCCCTGCAGACCATGGAGATGAGTGCTCGCTATCTTCAGTCATGTCACTTAAT
GCTAT

>gnl|SRA|SRR9643845.3.1 CB0WJANXX:4:1101:1438:2227 forward (Biological)
GTACAGACAGCCAGGCAGCTGCCCTCTCCTCCCCAGCCAAAGACTAGAGAGAGGTTCATT
TTCTAAATACGTGGAATCAGAAGCCCGACGGCTGAAGGTGGAACTGTTTGTACTGGAAAC
AGAAG

>gnl|SRA|SRR9643845.3.2 CB0WJANXX:4:1101:1438:2227 reverse (Biological)
CTCCTGGGATCTAGTCCTCATCAATTTGGCCTACAGGCTTTCACACAATGCTTCTGTTTC
CAGTACAAACAGTTCCACCTTCAGCCGTCGGGCTTCTTATTTCTCGTATTTATAAAATGG
AACTC

>gnl|SRA|SRR9643845.4.1 CB0WJANXX:4:1101:1739:2033 forward (Biological)
NGGCTCGCTGGCGTGGAGCCGGGCGTGGAATGCGAGTGCCTAGTGGGCCACTTTTGGTAA
GCAGAACTGGCGCTGCGGGATGAACCGAACGCCGGGTTAAGGCGCCCGATGCCGACGCTC
ATCAG

>gnl|SRA|SRR9643845.4.2 CB0WJANXX:4:1101:1739:2033 reverse (Biological)
CATCACACTAGTCGCTGCTGTCCTTATCAACCAACACCTTTTCTGGGGTCTGATGAGCGT
CGGCATCGGGCGCCTTAACCCGGCGTTCGGGTCATCCCGCAGCGCCAGTTCTGCTTATCC
AAAGT

>gnl|SRA|SRR9643845.6.1 CB0WJANXX:4:1101:1627:2069 forward (Biological)
CTTGAGCTCTGCAGTACCTCCTACATCTACACTTGATTTCCTTCCACTTTATCATTCTGT
TCTTCTCCAAGCAAGGCTCTAGTTGAGCTCCTGAGCTCGCTCTTCTTACCAGCACCCTCC
ACCTG

>gnl|SRA|SRR9643845.6.2 CB0WJANXX:4:1101:1627:2069 reverse (Biological)
TAGCTATGATGACAAAATTTTTAAAGTAAGTGAAAACACTCCCTATGATGGCAACCATGT
GGAGGGTGCTGGTAAGAAGAGCGAGCTCAGGAGCTCAACTAGAGCCTTGCTTGGAGAAGA
ACAGA

>gnl|SRA|SRR9643845.8.1 CB0WJANXX:4:1101:1696:2219 forward (Biological)
CATCACACTAGTCGCTGCCCCCGTCCACCTTTCTTCTCCGACCGACCGCCAGCGACGGCC
GGGTATGGGCCCGACGCTCCAGCGCCATCCATTTTCAGGGCTAGTTGATTCGGCAGGTGA
GTTGT

>gnl|SRA|SRR9643845.8.2 CB0WJANXX:4:1101:1696:2219 reverse (Biological)
AGACAGCAGGACGGTGGCCATGGAAGTCGGAATCCGCTAAGGAGTGTGTAACAACTCACC
TGCCGAATCAACTAGCCCTGAAAATGGATGGCGCTGGAGCGTCGGGCCCATACCCGGCCG
TCGCT

>gnl|SRA|SRR9643845.10.1 CB0WJANXX:4:1101:1663:2228 forward (Biological)
CCAAGTATATCAGATAAGATGCAAACAGGATCCCTCCATATCTCTGGGTTTCGAATCTGC
TGACTTAAGCAGCCACAGATGTGGAGGGCCATTACCCGAAAAGAATATGGTGGCAGTTCA
GATCT

>gnl|SRA|SRR9643845.10.2 CB0WJANXX:4:1101:1663:2228 reverse (Biological)
GTTGTCATAGGCTGCTCCCAAGATCTGAACTGCCACCATATTCTTTTCGGGTAATGGCCC
TCCACATCTGTGGCTGCTTAAGTCAGCAGATTCGAAACCCAGAGATATGGAGGGGTCCTG
TTTGG
```

Fig.8: the first 10 reads from SRR9643845. No significant level of telomere-like repeats were found.

An anomalous single-fragment amplicon from type strain Rabies Lyssavirus in SRR9644024

In Order to examine the property of SRR9644024, we performed a SERRATUS analysis of possible viral sequences in SRR9644024.



	Pan-genome coverage		
Match	Score	Identity	Reads
Polyomaviridae	96%		1040
AMR	84%		211
Hepadnaviridae	84%		1490
Parvoviridae	76%		2100
Circoviridae	28%		125
Rhabdoviridae	20%		2378
Alloherpesviridae	16%		12536
Anelloviridae	12%		82
Reoviridae	12%		88
Herpesviridae	12%		3701
Picornaviridae	8%		50
Coronaviridae	4%		6
Adenoviridae	4%		16
Filoviridae	4%		10
Astroviridae	0%		2
Poxviridae	0%		2
Genomoviridae	0%		4

Fig.9: the SERRATUS analysis of SRR9644024.

No RNA viral families exceeds pangenome coverage higher than 20%.

Furthermore, the major proportion of the reads, Rhabdoviridae, covers pangenome only 20%, despite the presence of over 2378 reads with sequencing depth of over 133x in the parts that were covered.

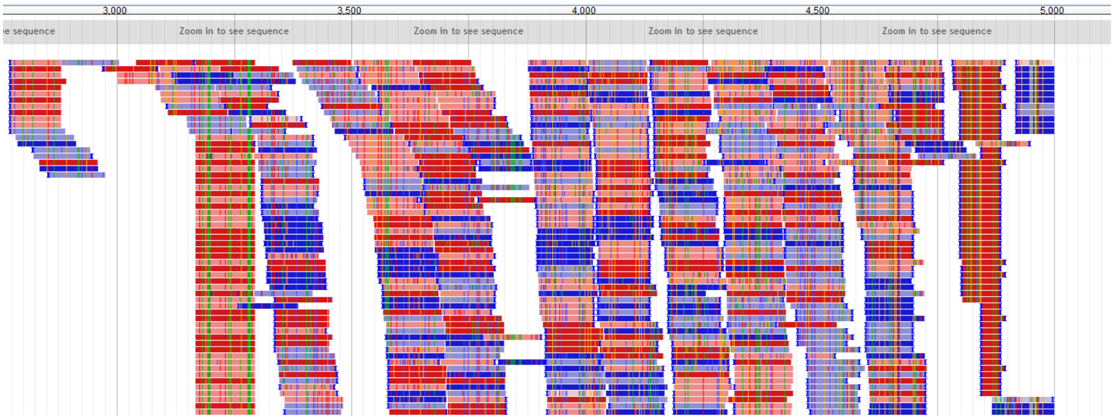
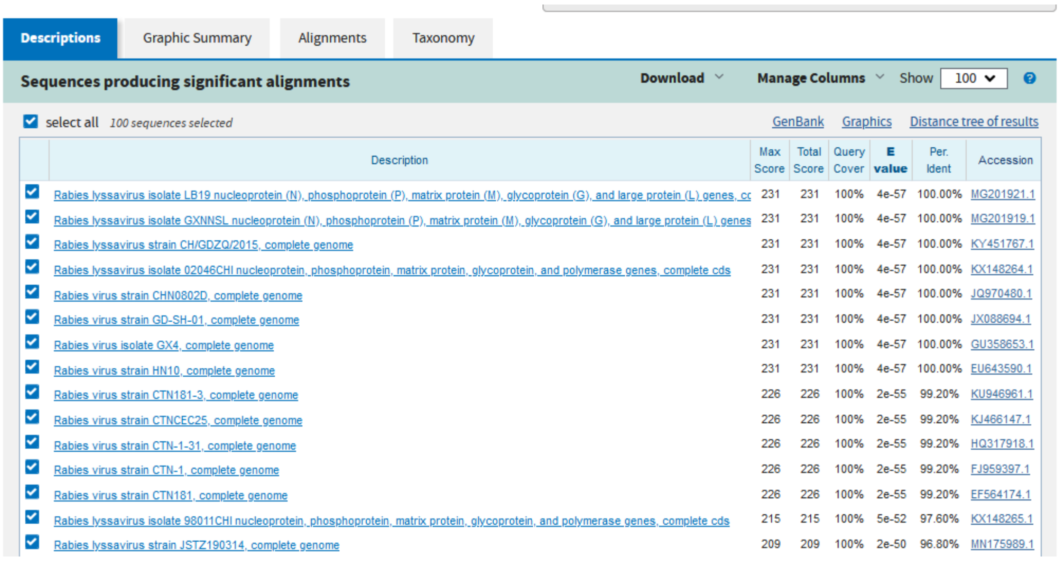


Fig.10: the single fragment of a Rabies Lyssavirus obtained from SRR9644024.

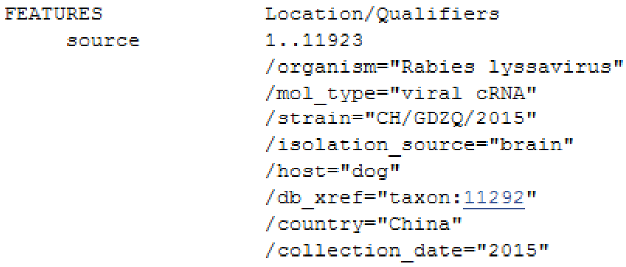
By BLASTing the obtained reads, the identity of the Lyssavirus was revealed to be the type strain CH/GDZQ/2015, which were isolated from the brains of dogs.



Sequences producing significant alignments

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

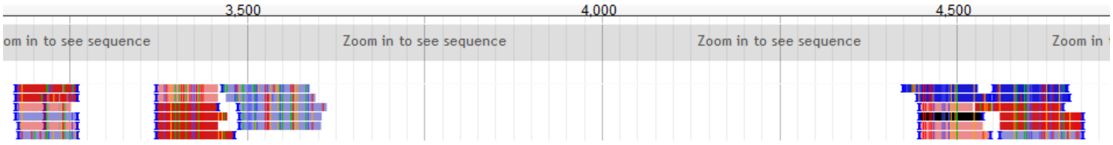
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus isolate LB19 nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large protein (L) genes, cc	231	231	100%	4e-57	100.00%	MG201921.1
☑	Rabies lyssavirus isolate GXNNSL nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large protein (L) genes	231	231	100%	4e-57	100.00%	MG201919.1
☑	Rabies lyssavirus strain CH/GDZQ/2015, complete genome	231	231	100%	4e-57	100.00%	KY451767.1
☑	Rabies lyssavirus isolate 02046CHI nucleoprotein, phosphoprotein, matrix protein, glycoprotein, and polymerase genes, complete cds	231	231	100%	4e-57	100.00%	KX148264.1
☑	Rabies virus strain CHN0802D, complete genome	231	231	100%	4e-57	100.00%	JQ970480.1
☑	Rabies virus strain GD-SH-01, complete genome	231	231	100%	4e-57	100.00%	JX088694.1
☑	Rabies virus isolate GX4, complete genome	231	231	100%	4e-57	100.00%	GU358653.1
☑	Rabies virus strain HN10, complete genome	231	231	100%	4e-57	100.00%	EU643590.1
☑	Rabies virus strain CTN181-3, complete genome	226	226	100%	2e-55	99.20%	KU946961.1
☑	Rabies virus strain CTNCEC25, complete genome	226	226	100%	2e-55	99.20%	KJ466147.1
☑	Rabies virus strain CTN-1-31, complete genome	226	226	100%	2e-55	99.20%	HQ317918.1
☑	Rabies virus strain CTN-1, complete genome	226	226	100%	2e-55	99.20%	FJ959397.1
☑	Rabies virus strain CTN181, complete genome	226	226	100%	2e-55	99.20%	EF564174.1
☑	Rabies lyssavirus isolate 98011CHI nucleoprotein, phosphoprotein, matrix protein, glycoprotein, and polymerase genes, complete cds	215	215	100%	5e-52	97.60%	KX148265.1
☑	Rabies lyssavirus strain JSTZ190314, complete genome	209	209	100%	2e-50	96.80%	MN175989.1

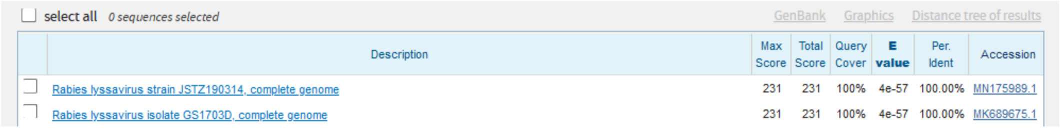


```
FEATURES             Location/Qualifiers
     source          1..11923
                     /organism="Rabies lyssavirus"
                     /mol_type="viral cRNA"
                     /strain="CH/GDZQ/2015"
                     /isolation_source="brain"
                     /host="dog"
                     /db_xref="taxon:11292"
                     /country="China"
                     /collection_date="2015"
```

Fig.11:The BLAST result and isolation host of the Rabies Lyssavirus reads.

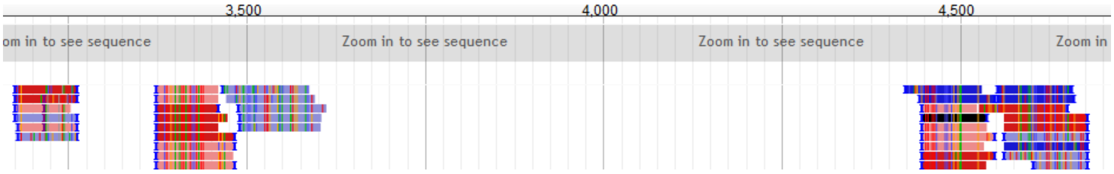
Despite being claimed as alignments to other Lyssavirus strains, all reads of Rhabdoviridae aligns to known type strains of Rabies Lyssavirus, indicating an origin as a single archived amplicon from a type culture.

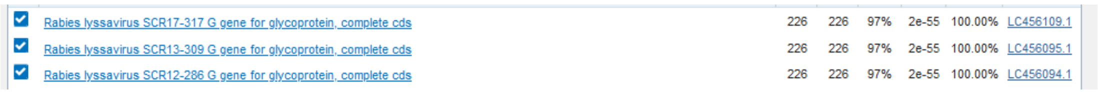




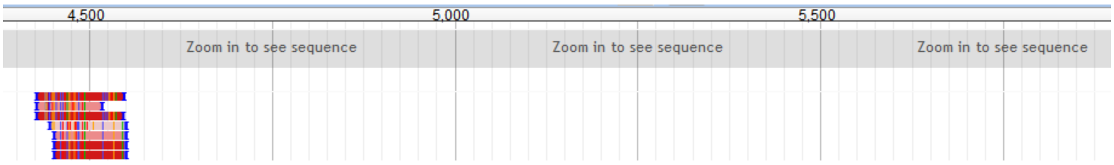
☐ select all *0 sequences selected* GenBank Graphics Distance tree of results

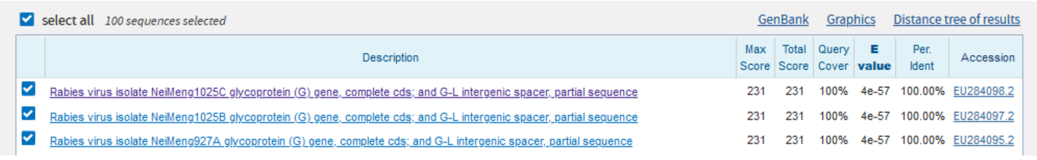
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☐	Rabies lyssavirus strain JSTZ190314, complete genome	231	231	100%	4e-57	100.00%	MN175989.1
☐	Rabies lyssavirus isolate GS1703D, complete genome	231	231	100%	4e-57	100.00%	MK689675.1





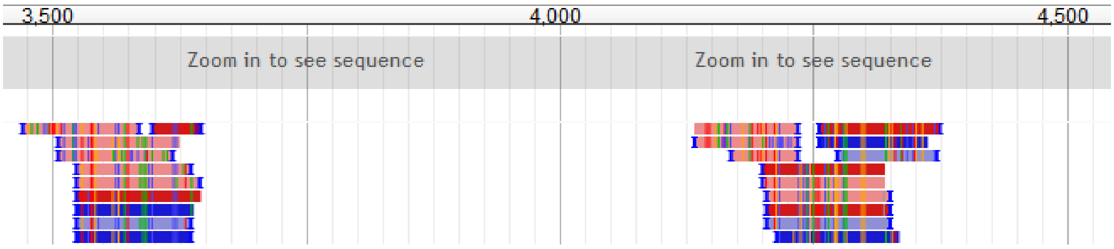
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus SCR17-317 G gene for glycoprotein, complete cds	226	226	97%	2e-55	100.00%	LC456109.1
☑	Rabies lyssavirus SCR13-309 G gene for glycoprotein, complete cds	226	226	97%	2e-55	100.00%	LC456095.1
☑	Rabies lyssavirus SCR12-286 G gene for glycoprotein, complete cds	226	226	97%	2e-55	100.00%	LC456094.1

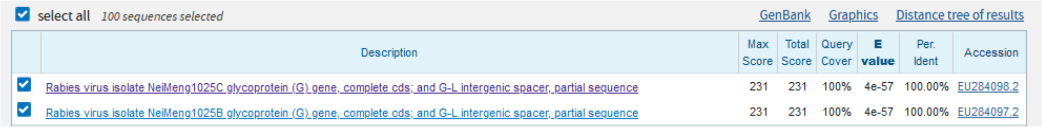




☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

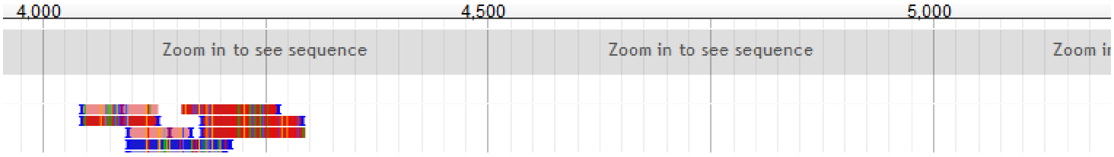
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies virus isolate NeiMeng1025C glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	231	231	100%	4e-57	100.00%	EU284098.2
☑	Rabies virus isolate NeiMeng1025B glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	231	231	100%	4e-57	100.00%	EU284097.2
☑	Rabies virus isolate NeiMeng927A glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	231	231	100%	4e-57	100.00%	EU284095.2

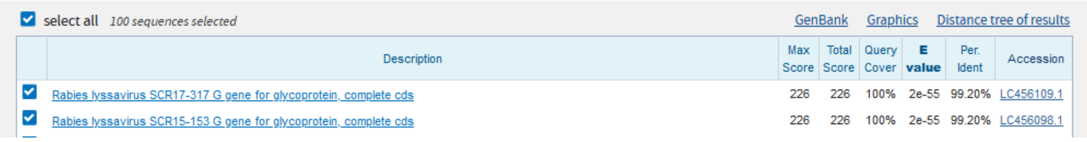




☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

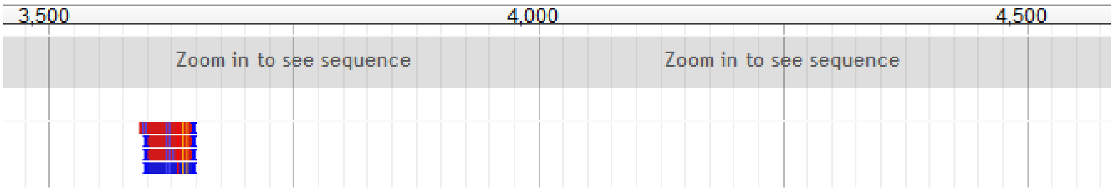
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies virus isolate NeiMeng1025C glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	231	231	100%	4e-57	100.00%	EU284098.2
☑	Rabies virus isolate NeiMeng1025B glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	231	231	100%	4e-57	100.00%	EU284097.2

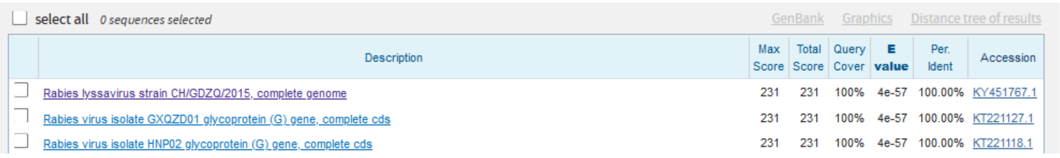




☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

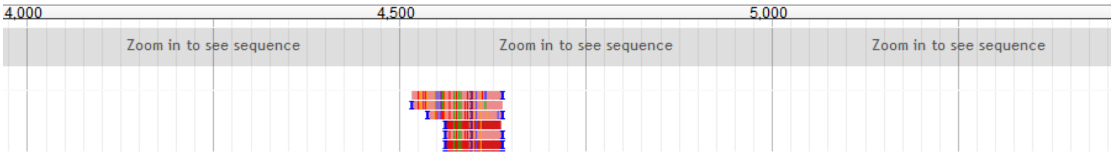
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus SCR17-317 G gene for glycoprotein, complete cds	226	226	100%	2e-55	99.20%	LC456109.1
☑	Rabies lyssavirus SCR15-153 G gene for glycoprotein, complete cds	226	226	100%	2e-55	99.20%	LC456098.1

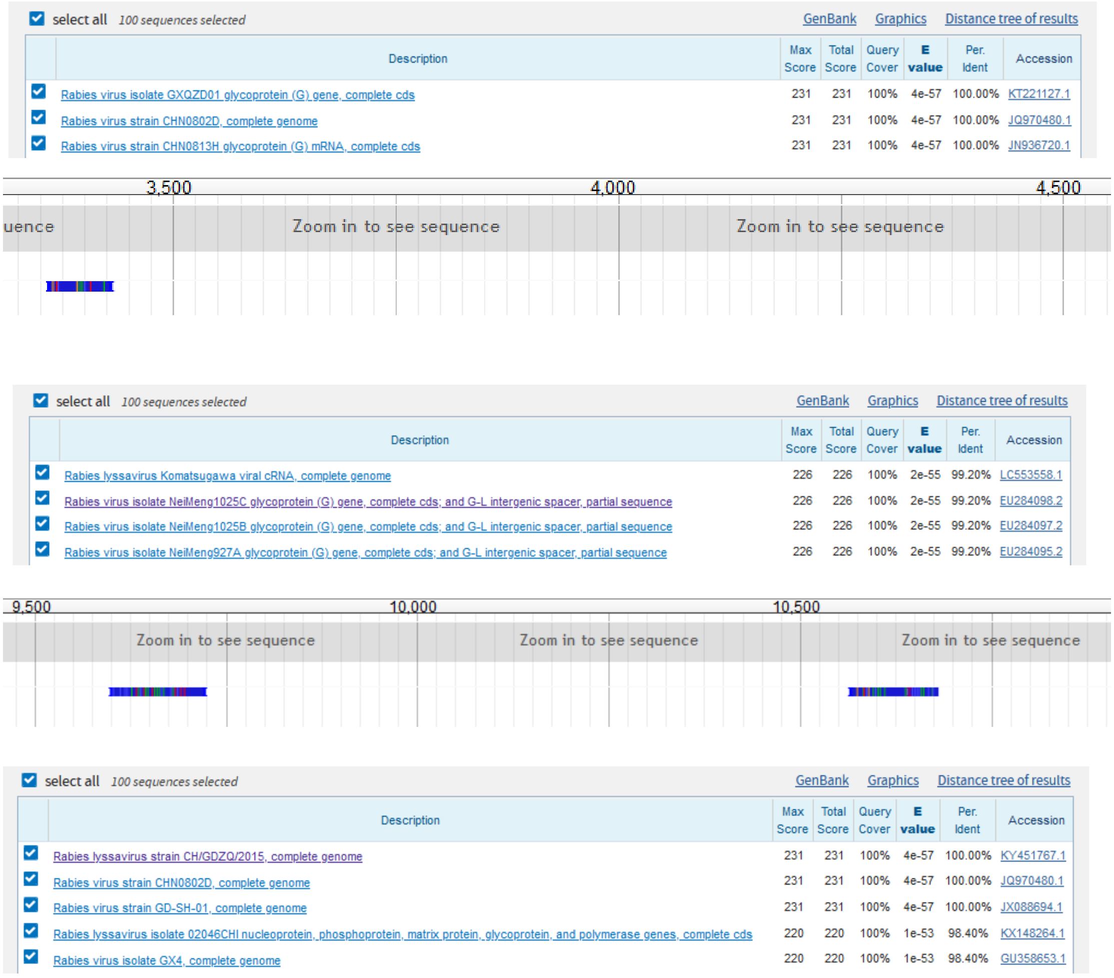




□ select all *0 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	Rabies lyssavirus strain CH/GDZQ/2015, complete genome	231	231	100%	4e-57	100.00%	KY451767.1
□	Rabies virus isolate GXQZD01 glycoprotein (G) gene, complete cds	231	231	100%	4e-57	100.00%	KT221127.1
□	Rabies virus isolate HNP02 glycoprotein (G) gene, complete cds	231	231	100%	4e-57	100.00%	KT221118.1





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

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies virus isolate GXQZD01 glycoprotein (G) gene, complete cds	231	231	100%	4e-57	100.00%	KT221127.1
☑	Rabies virus strain CHN0802D, complete genome	231	231	100%	4e-57	100.00%	JQ970480.1
☑	Rabies virus strain CHN0813H glycoprotein (G) mRNA, complete cds	231	231	100%	4e-57	100.00%	JN936720.1

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

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus Komatsugawa viral cRNA, complete genome	226	226	100%	2e-55	99.20%	LC553558.1
☑	Rabies virus isolate NeiMeng1025C glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	226	226	100%	2e-55	99.20%	EU284098.2
☑	Rabies virus isolate NeiMeng1025B glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	226	226	100%	2e-55	99.20%	EU284097.2
☑	Rabies virus isolate NeiMeng927A glycoprotein (G) gene, complete cds; and G-L intergenic spacer, partial sequence	226	226	100%	2e-55	99.20%	EU284095.2

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

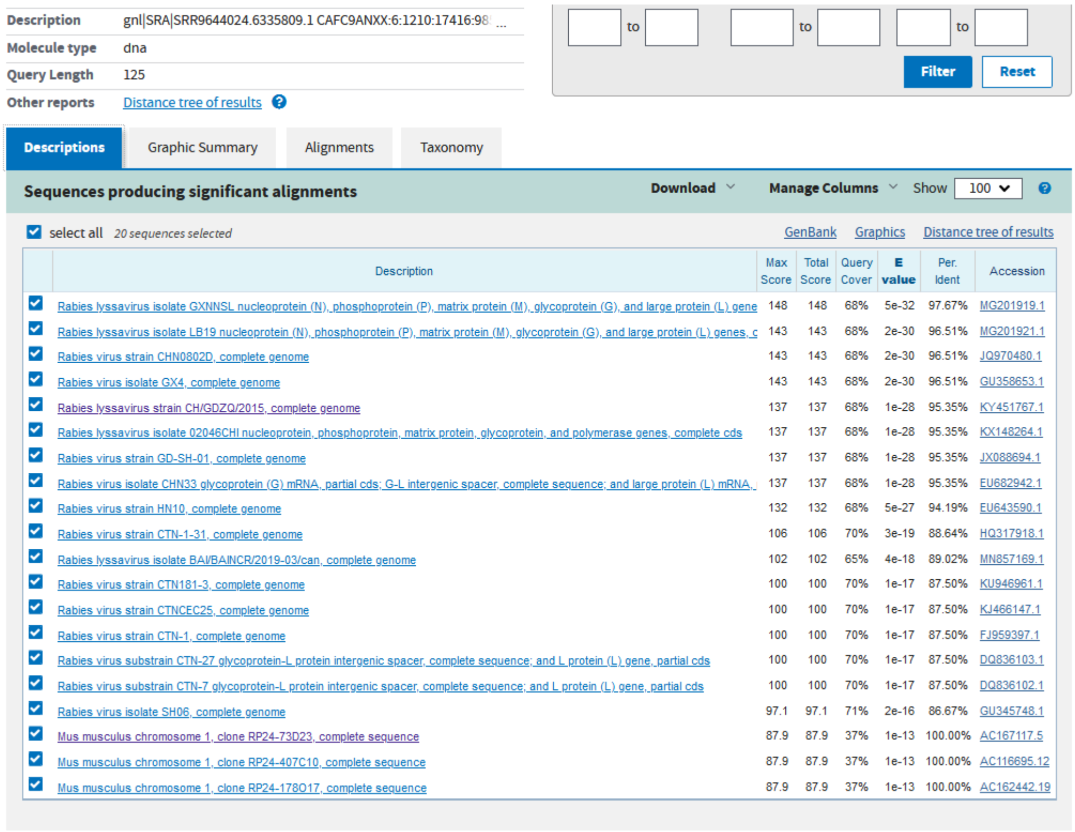
	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus strain CH/GDZQ/2015, complete genome	231	231	100%	4e-57	100.00%	KY451767.1
☑	Rabies virus strain CHN0802D, complete genome	231	231	100%	4e-57	100.00%	JQ970480.1
☑	Rabies virus strain GD-SH-01, complete genome	231	231	100%	4e-57	100.00%	JX088694.1
☑	Rabies lyssavirus isolate 02046CHI nucleoprotein, phosphoprotein, matrix protein, glycoprotein, and polymerase genes, complete cds	220	220	100%	1e-53	98.40%	KX148264.1
☑	Rabies virus isolate GX4, complete genome	220	220	100%	1e-53	98.40%	GU358653.1

Fig.12: reads claimed to be aligned to other Rhabdovirus genomes by SERRATUS. All came from the same amplicon of the G protein and a part of the M protein from Rabies virus type strains. We also obtained 2 aligned reads from the L protein of the same strain, which was the only reads that lands outside the anomalous amplicon.

The nature of the Rabies Lyssavirus reads as an amplicon isolated from Mus Musculus.

As Rabies Lyssavirus is a mononegavirus with a non-segmented genome, it is extremely improbable for a total nucleic acid preparation procedure to generate an extremely high coverage on one specific fragment of the viral genome yet did not cover any other part of the viral genome. In deed, the only other reads that was recovered from outside of this amplicon was 2 reads from the L protein of the exact same strain, which is most likely leftover templates from the PCR reaction.

In order to further characterize the nature of the anomalous amplicon-like reads, we BLASTed the reads that lands on the very end of the contig, which revealed that these reads were of a chimeric origin—DNA from Mus Musculus was found at the 3’-end of the Contig, while a highly conserved 22-mer sequence that lands in between the M and G region of most rabies Lyssavirus isolates (the region itself of which was Within the contig, rather than at the end of the contig), was found in the extreme 5’ end of the Contig.



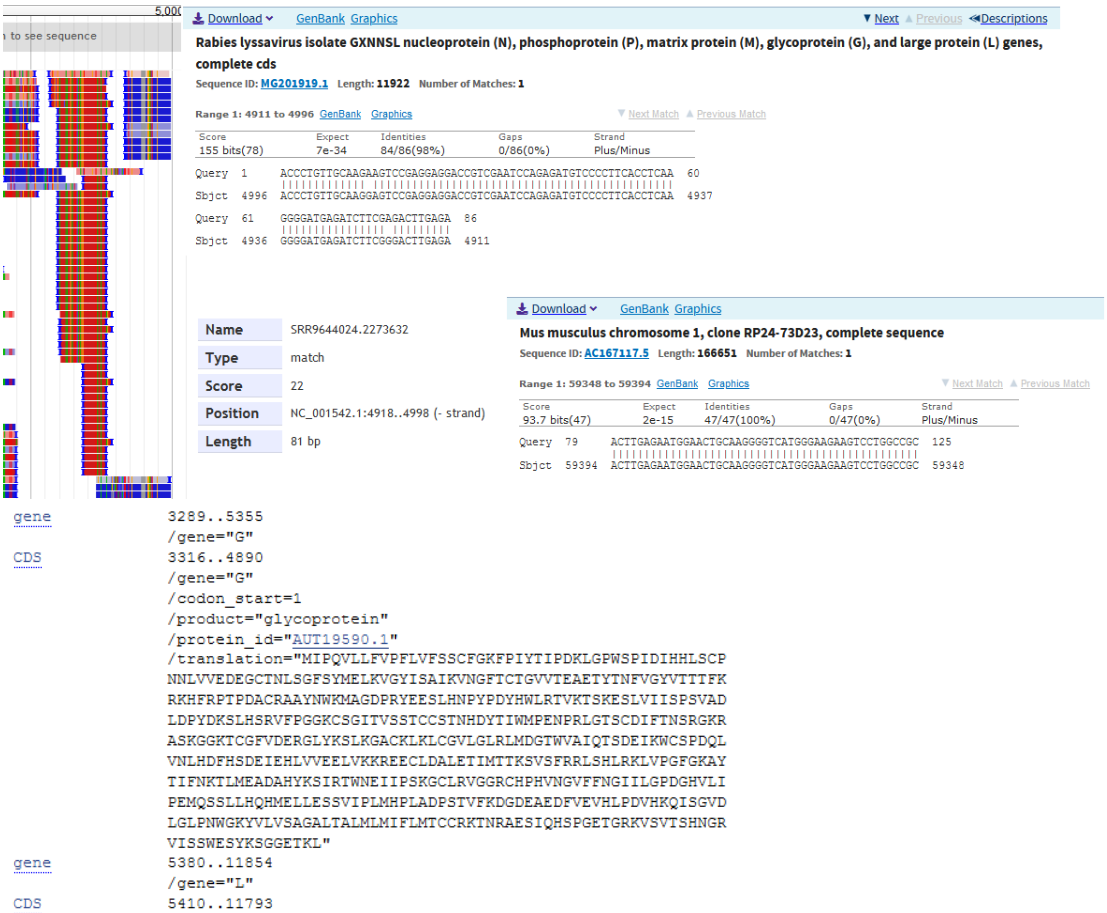
Description	gnl|SRA|SRR9644024.6335809.1 CAFC9ANXX:6:1210:17416:98 ...
Molecule type	dna
Query Length	125
Other reports	Distance tree of results

Sequences producing significant alignments

☑ select all *20 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Rabies lyssavirus isolate GXNNSL nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large protein (L) gene	148	148	68%	5e-32	97.67%	MG201919.1
☑	Rabies lyssavirus isolate LB19 nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large protein (L) genes, c	143	143	68%	2e-30	96.51%	MG201921.1
☑	Rabies virus strain CHN0802D, complete genome	143	143	68%	2e-30	96.51%	JQ970480.1
☑	Rabies virus isolate GX4, complete genome	143	143	68%	2e-30	96.51%	GU358653.1
☑	Rabies lyssavirus strain CH/GDZQ/2015, complete genome	137	137	68%	1e-28	95.35%	KY451767.1
☑	Rabies lyssavirus isolate 02046CHI nucleoprotein, phosphoprotein, matrix protein, glycoprotein, and polymerase genes, complete cds	137	137	68%	1e-28	95.35%	KX148264.1
☑	Rabies virus strain GD-SH-01, complete genome	137	137	68%	1e-28	95.35%	JX088694.1
☑	Rabies virus isolate CHN33 glycoprotein (G) mRNA, partial cds; G-L intergenic spacer, complete sequence; and large protein (L) mRNA,	137	137	68%	1e-28	95.35%	EU682942.1
☑	Rabies virus strain HN10, complete genome	132	132	68%	5e-27	94.19%	EU643590.1
☑	Rabies virus strain CTN-1-31, complete genome	106	106	70%	3e-19	88.64%	HQ317918.1
☑	Rabies lyssavirus isolate BAI/BAINCR/2019-03/can, complete genome	102	102	65%	4e-18	89.02%	MN857169.1
☑	Rabies virus strain CTN181-3, complete genome	100	100	70%	1e-17	87.50%	KU946961.1
☑	Rabies virus strain CTNCEC25, complete genome	100	100	70%	1e-17	87.50%	KJ466147.1
☑	Rabies virus strain CTN-1, complete genome	100	100	70%	1e-17	87.50%	FJ959397.1
☑	Rabies virus substrain CTN-27 glycoprotein-L protein intergenic spacer, complete sequence; and L protein (L) gene, partial cds	100	100	70%	1e-17	87.50%	DQ836103.1
☑	Rabies virus substrain CTN-7 glycoprotein-L protein intergenic spacer, complete sequence; and L protein (L) gene, partial cds	100	100	70%	1e-17	87.50%	DQ836102.1
☑	Rabies virus isolate SH06, complete genome	97.1	97.1	71%	2e-16	86.67%	GU345748.1
☑	Mus musculus chromosome 1, clone RP24-73D23, complete sequence	87.9	87.9	37%	1e-13	100.00%	AC167117.5
☑	Mus musculus chromosome 1, clone RP24-407C10, complete sequence	87.9	87.9	37%	1e-13	100.00%	AC116695.12
☑	Mus musculus chromosome 1, clone RP24-178O17, complete sequence	87.9	87.9	37%	1e-13	100.00%	AC162442.19

Fig.13a: the Mus Musculus DNA found at the extreme 3’ end of the Contig.



Rabies lyssavirus isolate GXNNSL nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large protein (L) genes, complete cds

Sequence ID: MG201919.1 Length: 11922 Number of Matches: 1

Range 1: 4911 to 4996 GenBank Graphics

```
Score          Expect   Identities     Gaps       Strand
155 bits(78)   7e-34    84/86(98%)     0/86(0%)   Plus/Minus

Query  1     ACCCTGTTGCAAGAAGTCCGAGGAGGACCGTCGAATCCAGAGATGTCCCCTTCACCTCAA  60
             ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4996  ACCCTGTTGCAAGGAGTCCGAGGAGGACCGTCGAATCCAGAGATGTCCCCTTCACCTCAA  4937

Query  61    GGGGATGAGATCTTCGAGACTTGAGA  86
             |||||||||||||||| |||||||||
Sbjct  4936  GGGGATGAGATCTTCGGGACTTGAGA  4911
```

Name	SRR9644024.2273632
Type	match
Score	22
Position	NC_001542.1:4918..4998 (- strand)
Length	81 bp

Mus musculus chromosome 1, clone RP24-73D23, complete sequence

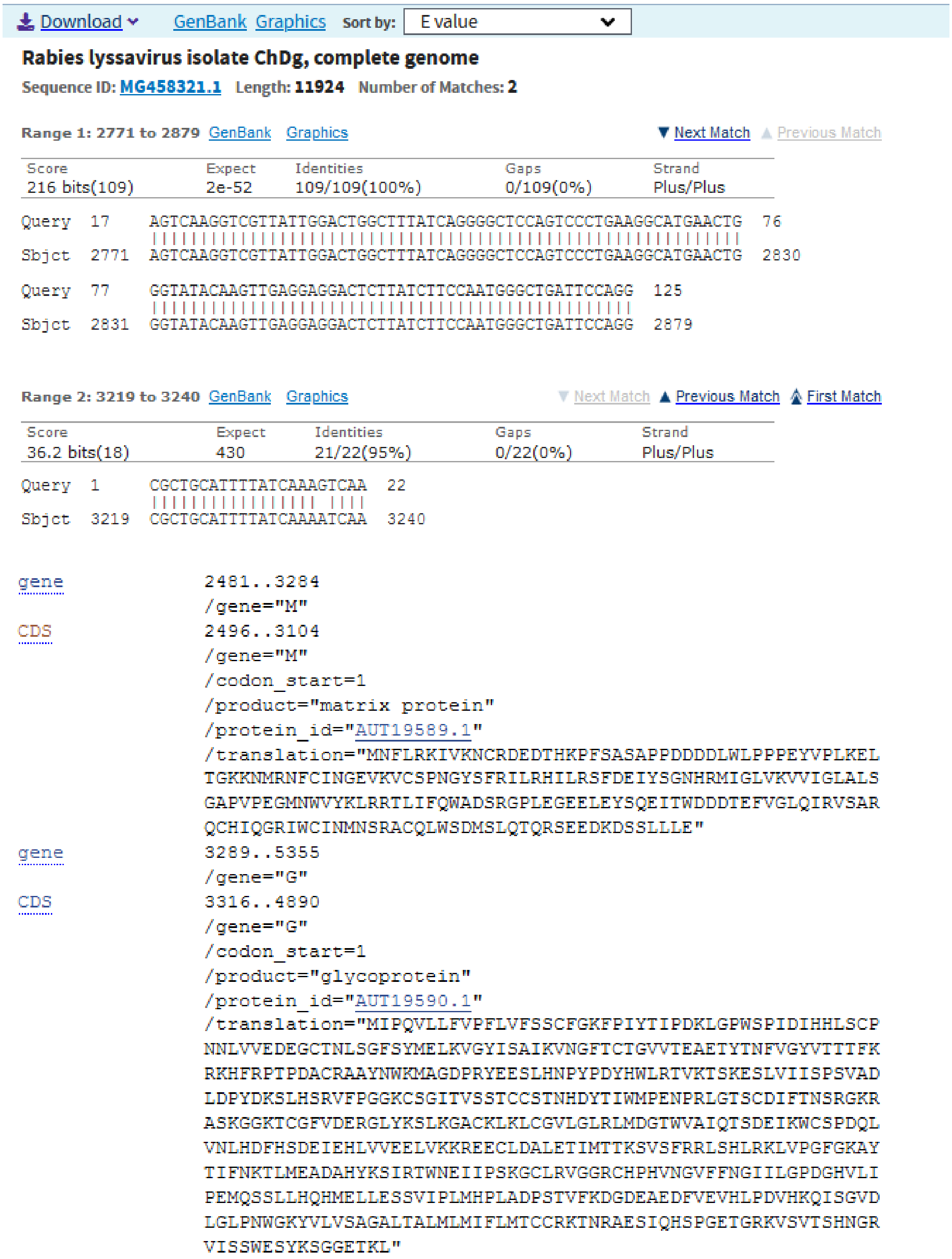
Sequence ID: AC167117.5 Length: 166651 Number of Matches: 1

Range 1: 59348 to 59394 GenBank Graphics

```
Score          Expect   Identities     Gaps       Strand
93.7 bits(47)  2e-15    47/47(100%)    0/47(0%)   Plus/Minus

Query  79     ACTTGAGAATGGAACTGCAAGGGGTCATGGGAAGAAGTCCTGGCCGC  125
              |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  59394  ACTTGAGAATGGAACTGCAAGGGGTCATGGGAAGAAGTCCTGGCCGC  59348
```

```
gene            3289..5355
                /gene="G"
CDS             3316..4890
                /gene="G"
                /codon_start=1
                /product="glycoprotein"
                /protein_id="AUT19590.1"
                /translation="MIPQVLLFVPFLVFSSCFGKFPIYTIPDKLGPWSPIDIHHLSCP
                NNLVVEDEGCTNLSGFSYMELKVGYISAIKVNGFTCTGVVTEAETYTNFVGYVTTTFK
                RKHFRPTPDACRAAYNWKMAGDPRYEESLHNPYPDYHWLRTVKTSKESLVIISPSVAD
                LDPYDKSLHSRVFPGGKCSGITVSSTCCSTNHDYTIWMPENPRLGTSCDIFTNSRGKR
                ASKGGKTCGFVDERGLYKSLKGACKLKLCGVLGLRLMDGTWVAIQTSDEIKWCSPDQL
                VNLHDFHSDEIEHLVVEELVKKREECLDALETIMTTKSVSFRRLSHLRKLVPGFGKAY
                TIFNKTLMEADAHYKSIRTWNEIIPSKGCLRVGGRCHPHVNGVFFNGIILGPDGHVLI
                PEMQSSLLHQHMELLESSVIPLMHPLADPSTVFKDGDEAEDFVEVHLPDVHKQISGVD
                LGLPNWGKYVLVSAGALTALMLMIFLMTCCRKTNRAESIQHSPGETGRKVSVTSHNGR
                VISSWESYKSGGETKL"
gene            5380..11854
                /gene="L"
CDS             5410..11793
```

Fig.13b: the 3’-end of the Contig. Notice that it lands right between the G gene and the L gene.



Rabies lyssavirus isolate ChDg, complete genome

Sequence ID: MG458321.1 Length: 11924 Number of Matches: 2

Range 1: 2771 to 2879 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
216 bits(109)	2e-52	109/109(100%)	0/109(0%)	Plus/Plus

```
Query  17     AGTCAAGGTCGTTATTGGACTGGCTTTATCAGGGGCTCCAGTCCCTGAAGGCATGAACTG  76
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2771   AGTCAAGGTCGTTATTGGACTGGCTTTATCAGGGGCTCCAGTCCCTGAAGGCATGAACTG  2830

Query  77     GGTATACAAGTTGAGGAGGACTCTTATCTTCCAATGGGCTGATTCCAGG  125
              |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2831   GGTATACAAGTTGAGGAGGACTCTTATCTTCCAATGGGCTGATTCCAGG  2879
```

Range 2: 3219 to 3240 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
36.2 bits(18)	430	21/22(95%)	0/22(0%)	Plus/Plus

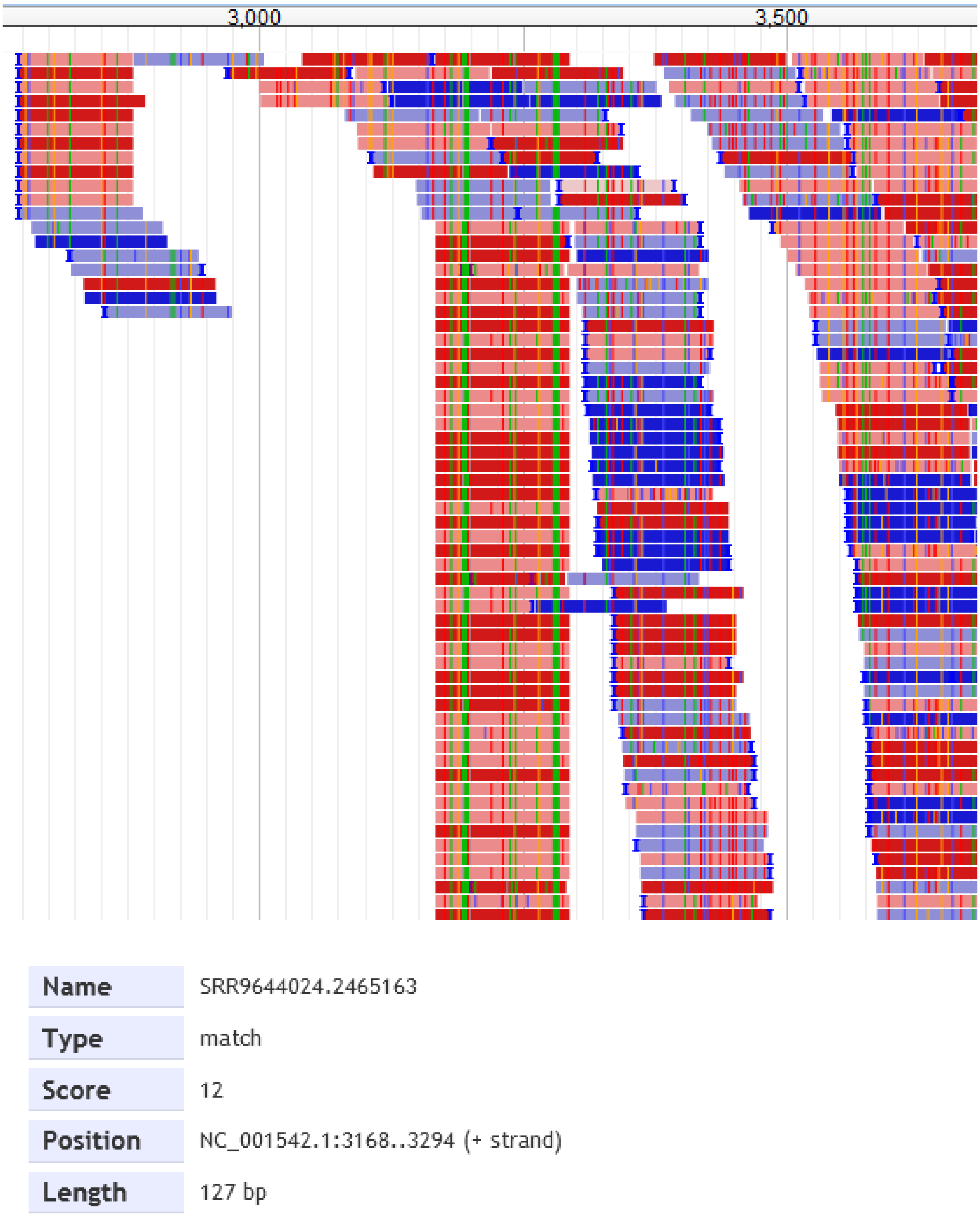
```
Query  1      CGCTGCATTTTATCAAAGTCAA  22
              |||||||||||||||| ||||
Sbjct  3219   CGCTGCATTTTATCAAAATCAA  3240
```

```
gene            2481..3284
                /gene="M"
CDS             2496..3104
                /gene="M"
                /codon_start=1
                /product="matrix protein"
                /protein_id="AUT19589.1"
                /translation="MNFLRKIVKNCRDEDTHKPFSASAPPDDDDLWLPPPEYVPLKEL
                TGKKNMRNFCINGEVKVCSPNGYSFRILRHILRSFDEIYSGNHRMIGLVKVVIGLALS
                GAPVPEGMNWVYKLRRTLIFQWADSRGPLEGEELEYSQEITWDDDTEFVGLQIRVSAR
                QCHIQGRIWCINMNSRACQLWSDMSLQTQRSEEDKDSSLLLE"
gene            3289..5355
                /gene="G"
CDS             3316..4890
                /gene="G"
                /codon_start=1
                /product="glycoprotein"
                /protein_id="AUT19590.1"
                /translation="MIPQVLLFVPFLVFSSCFGKFPIYTIPDKLGPWSPIDIHHLSCP
                NNLVVEDEGCTNLSGFSYMELKVGYISAIKVNGFTCTGVVTEAETYTNFVGYVTTTFK
                RKHFRPTPDACRAAYNWKMAGDPRYEESLHNPYPDYHWLRTVKTSKESLVIISPSVAD
                LDPYDKSLHSRVFPGGKCSGITVSSTCCSTNHDYTIWMPENPRLGTSCDIFTNSRGKR
                ASKGGKTCGFVDERGLYKSLKGACKLKLCGVLGLRLMDGTWVAIQTSDEIKWCSPDQL
                VNLHDFHSDEIEHLVVEELVKKREECLDALETIMTTKSVSFRRLSHLRKLVPGFGKAY
                TIFNKTLMEADAHYKSIRTWNEIIPSKGCLRVGGRCHPHVNGVFFNGIILGPDGHVLI
                PEMQSSLLHQHMELLESSVIPLMHPLADPSTVFKDGDEAEDFVEVHLPDVHKQISGVD
                LGLPNWGKYVLVSAGALTALMLMIFLMTCCRKTNRAESIQHSPGETGRKVSVTSHNGR
                VISSWESYKSGGETKL"
```

Fig.14: the misplaced 22-mer found at the extreme 5’-end of the Contig.

The position of such an 22-mer lands exactly where a primer for the amplification of the 5’-end of the G gene would be located, and is likely a product of mispriming of the PCR template.

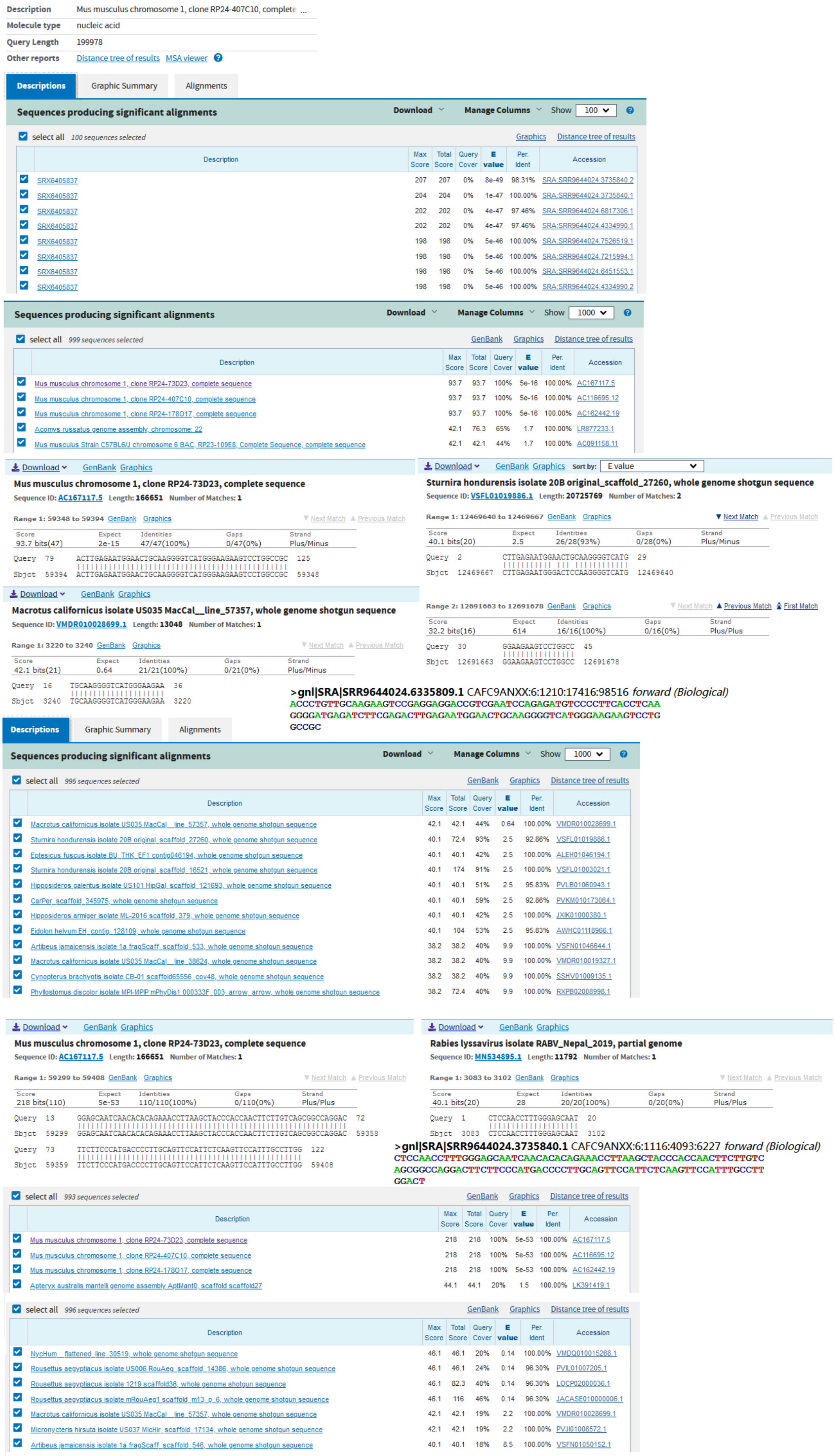
In deed, we discovered that the vast majority of the reads begins at position 3168, which a primer for amplifying the G protein would have been located at.



Name	SRR9644024.2465163
Type	match
Score	12
Position	NC_001542.1:3168..3294 (+ strand)
Length	127 bp

Fig.15: the beginning of the vast majority of the reads for Rabies Lyssavirus lands at position 3168.

These properties, including the fact that the extreme 5'-end and 3'-end sequence being exactly flanking the G protein, alongside with the presence of mispriming products containing Mus Musculus DNA, of which were not found in bats (WGS with 100 databases currently on NCBI, Chiroptera, txid: 9397), point toward the Rabies Lyssavirus being a PCR clone derived from Mus Musculus. It therefore constitutes a fraudulent sample material, which is likely introduced into SRR9644024 from the pooling process.



Description	Mus musculus chromosome 1, clone RP24-407C10, complete ...
Molecule type	nucleic acid
Query Length	199978
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
☑	SRX6405837	207	207	0%	8e-49	98.31%	SRA:SRR9644024.3735840.2
☑	SRX6405837	204	204	0%	1e-47	100.00%	SRA:SRR9644024.3735840.1
☑	SRX6405837	202	202	0%	4e-47	97.46%	SRA:SRR9644024.6817306.1
☑	SRX6405837	202	202	0%	4e-47	97.46%	SRA:SRR9644024.4334990.1
☑	SRX6405837	198	198	0%	5e-46	100.00%	SRA:SRR9644024.7526519.1
☑	SRX6405837	198	198	0%	5e-46	100.00%	SRA:SRR9644024.7215994.1
☑	SRX6405837	198	198	0%	5e-46	100.00%	SRA:SRR9644024.6451553.1
☑	SRX6405837	198	198	0%	5e-46	100.00%	SRA:SRR9644024.4334990.2

Sequences producing significant alignments

☑ select all *999 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Mus musculus chromosome 1, clone RP24-73D23, complete sequence	93.7	93.7	100%	5e-16	100.00%	AC167117.5
☑	Mus musculus chromosome 1, clone RP24-407C10, complete sequence	93.7	93.7	100%	5e-16	100.00%	AC116695.12
☑	Mus musculus chromosome 1, clone RP24-178O17, complete sequence	93.7	93.7	100%	5e-16	100.00%	AC162442.19
☑	Acomys russatus genome assembly, chromosome: 22	42.1	76.3	65%	1.7	100.00%	LR877233.1
☑	Mus musculus Strain C57BL6/J chromosome 6 BAC, RP23-109E8, Complete Sequence, complete sequence	42.1	42.1	44%	1.7	100.00%	AC091158.11

Mus musculus chromosome 1, clone RP24-73D23, complete sequence

Sequence ID: AC167117.5 Length: 166651 Number of Matches: 1

Range 1: 59348 to 59394 GenBank Graphics

```
Score          Expect    Identities     Gaps       Strand
93.7 bits(47)  2e-15     47/47(100%)    0/47(0%)   Plus/Minus

Query  79     ACTTGAGAATGGAACTGCAAGGGGTCATGGGAAGAAGTCCTGGCCGC  125
              |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  59394  ACTTGAGAATGGAACTGCAAGGGGTCATGGGAAGAAGTCCTGGCCGC  59348
```

Macrotus californicus isolate US035 MacCal__line_57357, whole genome shotgun sequence

Sequence ID: VMDR010028699.1 Length: 13048 Number of Matches: 1

Range 1: 3220 to 3240 GenBank Graphics

```
Score          Expect    Identities     Gaps       Strand
42.1 bits(21)  0.64      21/21(100%)    0/21(0%)   Plus/Minus

Query  16    TGCAAGGGGTCATGGGAAGAA  36
             |||||||||||||||||||||
Sbjct  3240  TGCAAGGGGTCATGGGAAGAA  3220
```

Sturnira hondurensis isolate 20B original_scaffold_27260, whole genome shotgun sequence

Sequence ID: VSFL01019886.1 Length: 20725769 Number of Matches: 2

Range 1: 12469640 to 12469667 GenBank Graphics

```
Score          Expect    Identities     Gaps       Strand
40.1 bits(20)  2.5       26/28(93%)     0/28(0%)   Plus/Minus

Query  2         CTTGAGAATGGAACTGCAAGGGGTCATG  29
                 ||||||||||| |||  |||||||||||
Sbjct  12469667  CTTGAGAATGGGACTCCAAGGGGTCATG  12469640
```

Range 2: 12691663 to 12691678 GenBank Graphics

```
Score          Expect    Identities     Gaps       Strand
32.2 bits(16)  614       16/16(100%)    0/16(0%)   Plus/Plus

Query  30        GGAAGAAGTCCTGGCC  45
                 ||||||||||||||||
Sbjct  12691663  GGAAGAAGTCCTGGCC  12691678
```

>gnl|SRA|SRR9644024.6335809.1 CAFC9ANXX:6:1210:17416:98516 *forward (Biological)*

```
ACCCTGTTGCAAGAAGTCCGAGGAGGACCGTCGAATCCAGAGATGTCCCCTTCACCTCAA
GGGGATGAGATCTTCGAGACTTGAGAATGGAACTGCAAGGGGTCATGGGAAGAAGTCCTG
GCCGC
```

Sequences producing significant alignments

☑ select all *995 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Macrotus californicus isolate US035 MacCal__line_57357, whole genome shotgun sequence	42.1	42.1	44%	0.64	100.00%	VMDR010028699.1
☑	Sturnira hondurensis isolate 20B original_scaffold_27260, whole genome shotgun sequence	40.1	72.4	93%	2.5	92.86%	VSFL01019886.1
☑	Eptesicus fuscus isolate BU_THK_EF1 contig046194, whole genome shotgun sequence	40.1	40.1	42%	2.5	100.00%	ALEH01046194.1
☑	Sturnira hondurensis isolate 20B original_scaffold_16521, whole genome shotgun sequence	40.1	174	91%	2.5	100.00%	VSFL01003021.1
☑	Hipposideros galeritus isolate US101 HipGal_scaffold_121693, whole genome shotgun sequence	40.1	40.1	51%	2.5	95.83%	PVLB01060943.1
☑	CarPer_scaffold_345975, whole genome shotgun sequence	40.1	40.1	59%	2.5	92.86%	PVKM010173064.1
☑	Hipposideros armiger isolate ML-2016 scaffold_379, whole genome shotgun sequence	40.1	40.1	42%	2.5	100.00%	JXIK01000380.1
☑	Eidolon helvum EH_contig_128109, whole genome shotgun sequence	40.1	104	53%	2.5	95.83%	AWHC01118966.1
☑	Artibeus jamaicensis isolate 1a fragScaff_scaffold_533, whole genome shotgun sequence	38.2	38.2	40%	9.9	100.00%	VSFN01046644.1
☑	Macrotus californicus isolate US035 MacCal__line_38624, whole genome shotgun sequence	38.2	38.2	40%	9.9	100.00%	VMDR010019327.1
☑	Cynopterus brachyotis isolate CB-01 scaffold65556_cov48, whole genome shotgun sequence	38.2	38.2	40%	9.9	100.00%	SSHV01009135.1
☑	Phyllostomus discolor isolate MPI-MPIP mPhyDis1 000333F_003_arrow_arrow, whole genome shotgun sequence	38.2	72.4	40%	9.9	100.00%	RXPB02008998.1

Mus musculus chromosome 1, clone RP24-73D23, complete sequence

Sequence ID: AC167117.5 Length: 166651 Number of Matches: 1

Range 1: 59299 to 59408 GenBank Graphics

```
Score           Expect    Identities      Gaps        Strand
218 bits(110)   5e-53     110/110(100%)   0/110(0%)   Plus/Plus

Query  13     GGAGCAATCAACACACAGAAACCTTAAGCTACCCACCAACTTCTTGTCAGCGGCCAGGAC  72
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  59299  GGAGCAATCAACACACAGAAACCTTAAGCTACCCACCAACTTCTTGTCAGCGGCCAGGAC  59358

Query  73     TTCTTCCCATGACCCCTTGCAGTTCCATTCTCAAGTTCCATTTGCCTTGG  122
              ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  59359  TTCTTCCCATGACCCCTTGCAGTTCCATTCTCAAGTTCCATTTGCCTTGG  59408
```

Rabies lyssavirus isolate RABV_Nepal_2019, partial genome

Sequence ID: MN534895.1 Length: 11792 Number of Matches: 1

Range 1: 3083 to 3102 GenBank Graphics

```
Score          Expect    Identities     Gaps       Strand
40.1 bits(20)  28        20/20(100%)    0/20(0%)   Plus/Plus

Query  1     CTCCAACCTTTGGGAGCAAT  20
             ||||||||||||||||||||
Sbjct  3083  CTCCAACCTTTGGGAGCAAT  3102
```

>gnl|SRA|SRR9644024.3735840.1 CAFC9ANXX:6:1116:4093:6227 *forward (Biological)*

```
CTCCAACCTTTGGGAGCAATCAACACACAGAAACCTTAAGCTACCCACCAACTTCTTGTC
AGCGGCCAGGACTTCTTCCCATGACCCCTTGCAGTTCCATTCTCAAGTTCCATTTGCCTT
GGACT
```

☑ select all *993 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Mus musculus chromosome 1, clone RP24-73D23, complete sequence	218	218	100%	5e-53	100.00%	AC167117.5
☑	Mus musculus chromosome 1, clone RP24-407C10, complete sequence	218	218	100%	5e-53	100.00%	AC116695.12
☑	Mus musculus chromosome 1, clone RP24-178O17, complete sequence	218	218	100%	5e-53	100.00%	AC162442.19
☑	Apteryx australis mantelli genome assembly AptMant0, scaffold scaffold27	44.1	44.1	20%	1.5	100.00%	LK391419.1

☑ select all *996 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	NycHum__flattened_line_30519, whole genome shotgun sequence	46.1	46.1	20%	0.14	100.00%	VMDQ010015268.1
☑	Rousettus aegyptiacus isolate US006 RouAeg_scaffold_14386, whole genome shotgun sequence	46.1	46.1	24%	0.14	96.30%	PVIL01007205.1
☑	Rousettus aegyptiacus isolate 1219 scaffold36, whole genome shotgun sequence	46.1	82.3	40%	0.14	96.30%	LOCP02000036.1
☑	Rousettus aegyptiacus isolate mRouAeg1 scaffold_m13_p_6, whole genome shotgun sequence	46.1	116	46%	0.14	96.30%	JACASE010000006.1
☑	Macrotus californicus isolate US035 MacCal__line_57357, whole genome shotgun sequence	42.1	42.1	19%	2.2	100.00%	VMDR010028699.1
☑	Micronycteris hirsuta isolate US037 MicHir_scaffold_17134, whole genome shotgun sequence	42.1	42.1	19%	2.2	100.00%	PVJI01008572.1
☑	Artibeus jamaicensis isolate 1a fragScaff_scaffold_546, whole genome shotgun sequence	40.1	40.1	18%	8.5	100.00%	VSFN01050152.1

Fig.16: Mus Musculus DNA found fused to the 3’ end of Rabies Lyssavirus, which were distinctly not bat or human in origin.

Presence of mispriming products from virus-specific primers in SRR9644024

We obtained 50 reads matching PicoRNAvirales from SRR9644024 covering 8% pangenome. However, reads matching PicoRNAvirales does not form non-duplicate contiguous sequences that can generate meaningful assemblies.

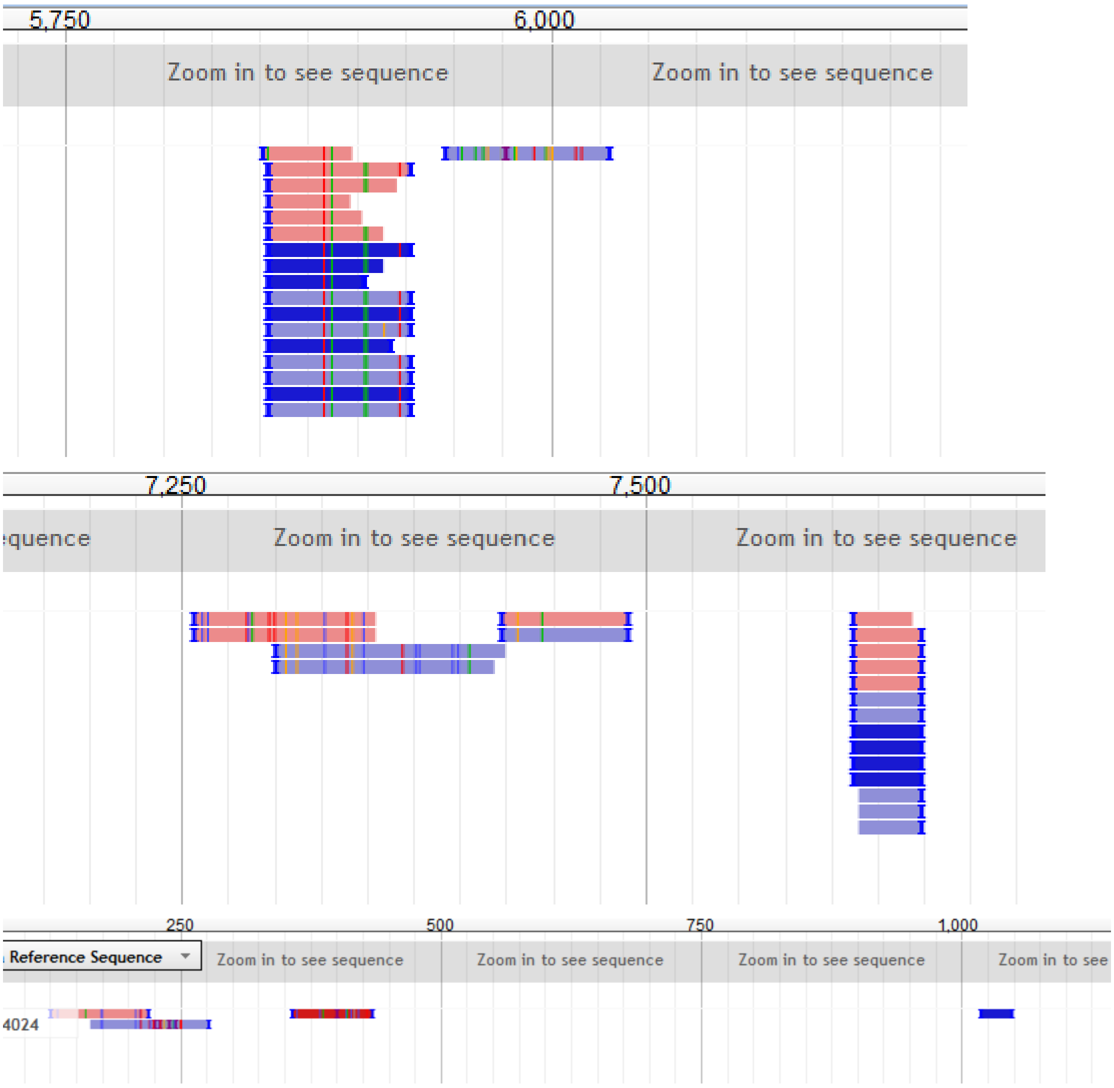
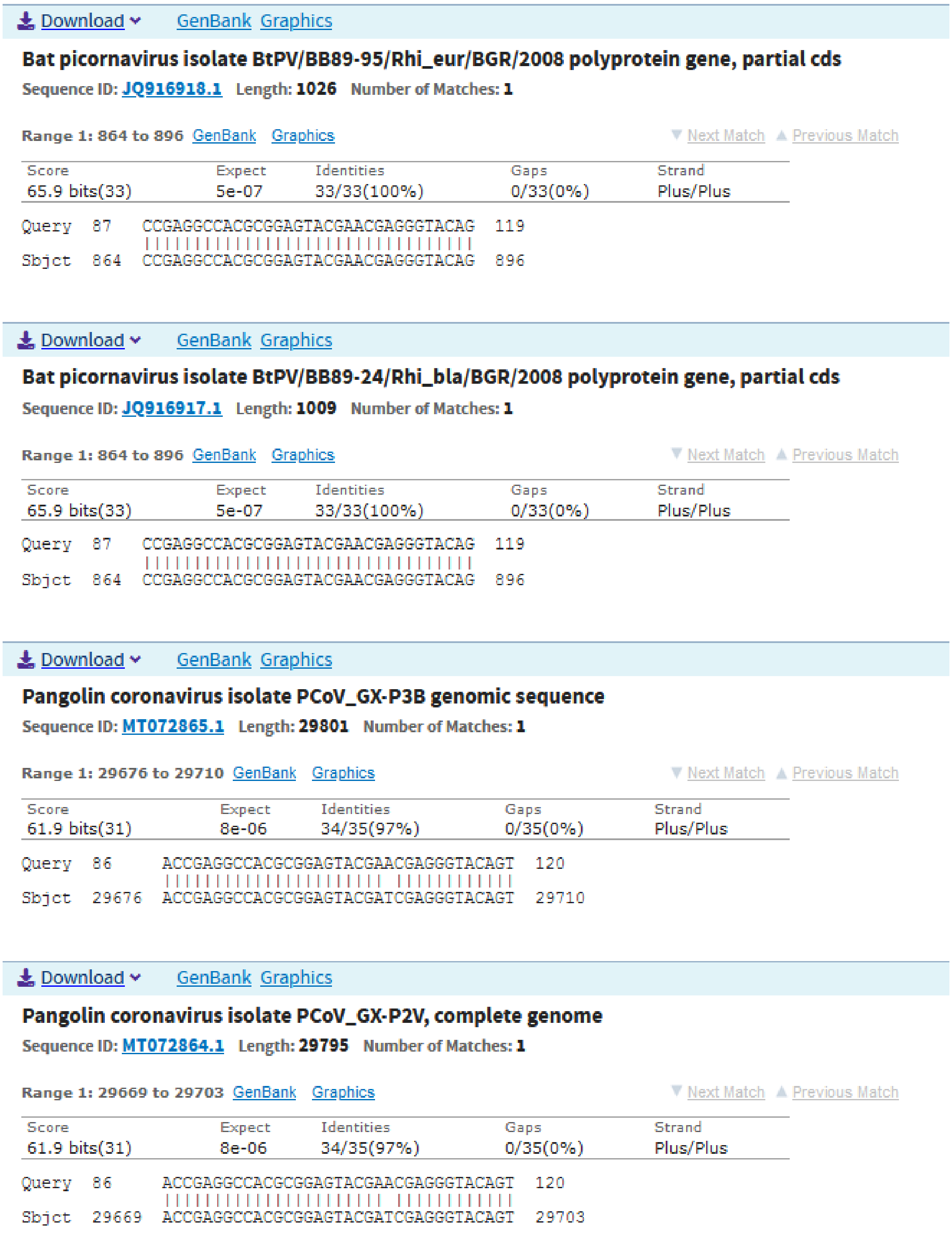


Fig.17: Reads matched to picoRNAvirales from SRR9644024.

Importantly, most of the reads from picoRNAvirales came from a partial match at the extreme 3’ end of the genome, which corresponds to a common 35-mer fond in the 3’-UTR of a diverse range of picoRNAviruses and coronaviruses. It seems to match best to a S2m motif, and despite extensive searching, we could not find any match to the regions flanking the 5’-end of this motif, suggesting it is likely the result of a mispriming product from a universal Pisunivirus primer to either random PCR ligation products or DNA contamination.



Bat picornavirus isolate BtPV/BB89-95/Rhi_eur/BGR/2008 polyprotein gene, partial cds

Sequence ID: JQ916918.1 Length: 1026 Number of Matches: 1

Range 1: 864 to 896 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
65.9 bits(33)	5e-07	33/33(100%)	0/33(0%)	Plus/Plus

```
Query  87   CCGAGGCCACGCGGAGTACGAACGAGGGTACAG  119
            |||||||||||||||||||||||||||||||||
Sbjct  864  CCGAGGCCACGCGGAGTACGAACGAGGGTACAG  896
```

Bat picornavirus isolate BtPV/BB89-24/Rhi_bla/BGR/2008 polyprotein gene, partial cds

Sequence ID: JQ916917.1 Length: 1009 Number of Matches: 1

Range 1: 864 to 896 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
65.9 bits(33)	5e-07	33/33(100%)	0/33(0%)	Plus/Plus

```
Query  87   CCGAGGCCACGCGGAGTACGAACGAGGGTACAG  119
            |||||||||||||||||||||||||||||||||
Sbjct  864  CCGAGGCCACGCGGAGTACGAACGAGGGTACAG  896
```

Pangolin coronavirus isolate PCoV_GX-P3B genomic sequence

Sequence ID: MT072865.1 Length: 29801 Number of Matches: 1

Range 1: 29676 to 29710 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
61.9 bits(31)	8e-06	34/35(97%)	0/35(0%)	Plus/Plus

```
Query  86     ACCGAGGCCACGCGGAGTACGAACGAGGGTACAGT  120
              |||||||||||||||||||||| ||||||||||||
Sbjct  29676  ACCGAGGCCACGCGGAGTACGATCGAGGGTACAGT  29710
```

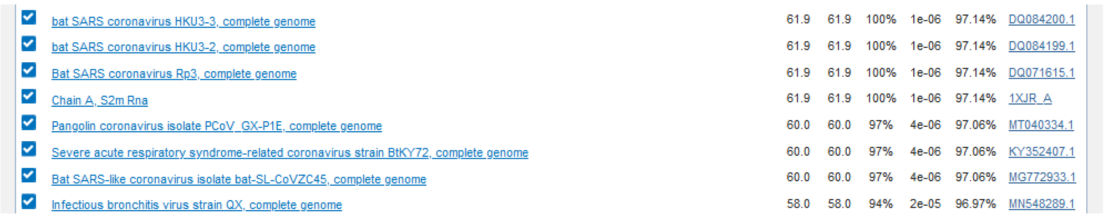
Pangolin coronavirus isolate PCoV_GX-P2V, complete genome

Sequence ID: MT072864.1 Length: 29795 Number of Matches: 1

Range 1: 29669 to 29703 GenBank Graphics

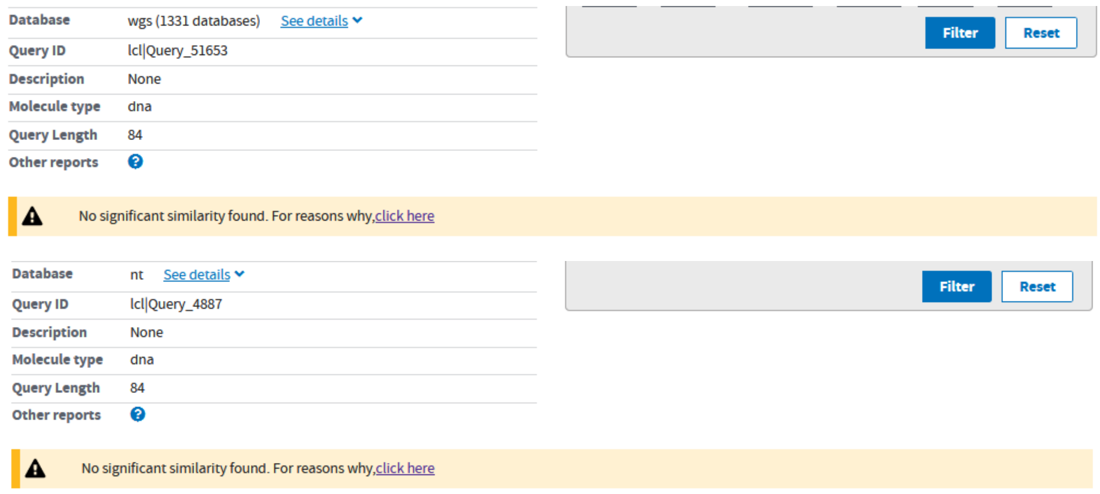
Score	Expect	Identities	Gaps	Strand
61.9 bits(31)	8e-06	34/35(97%)	0/35(0%)	Plus/Plus

```
Query  86     ACCGAGGCCACGCGGAGTACGAACGAGGGTACAGT  120
              |||||||||||||||||||||| ||||||||||||
Sbjct  29669  ACCGAGGCCACGCGGAGTACGATCGAGGGTACAGT  29703
```



☑	bat SARS coronavirus HKU3-3, complete genome	61.9	61.9	100%	1e-06	97.14%	DQ084200.1
☑	bat SARS coronavirus HKU3-2, complete genome	61.9	61.9	100%	1e-06	97.14%	DQ084199.1
☑	Bat SARS coronavirus Rp3, complete genome	61.9	61.9	100%	1e-06	97.14%	DQ071615.1
☑	Chain A, S2m Rna	61.9	61.9	100%	1e-06	97.14%	1XJR_A
☑	Pangolin coronavirus isolate PCoV_GX-P1E, complete genome	60.0	60.0	97%	4e-06	97.06%	MT040334.1
☑	Severe acute respiratory syndrome-related coronavirus strain BtKY72, complete genome	60.0	60.0	97%	4e-06	97.06%	KY352407.1
☑	Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome	60.0	60.0	97%	4e-06	97.06%	MG772933.1
☑	Infectious bronchitis virus strain QX, complete genome	58.0	58.0	94%	2e-05	96.97%	MN548289.1

Fig.18a: S2m motif found in SRR9644024 matching many diverse picoRNAviruses and Coronaviruses.



Database	wgs (1331 databases) See details
Query ID	lcl|Query_51653
Description	None
Molecule type	dna
Query Length	84
Other reports	❓

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

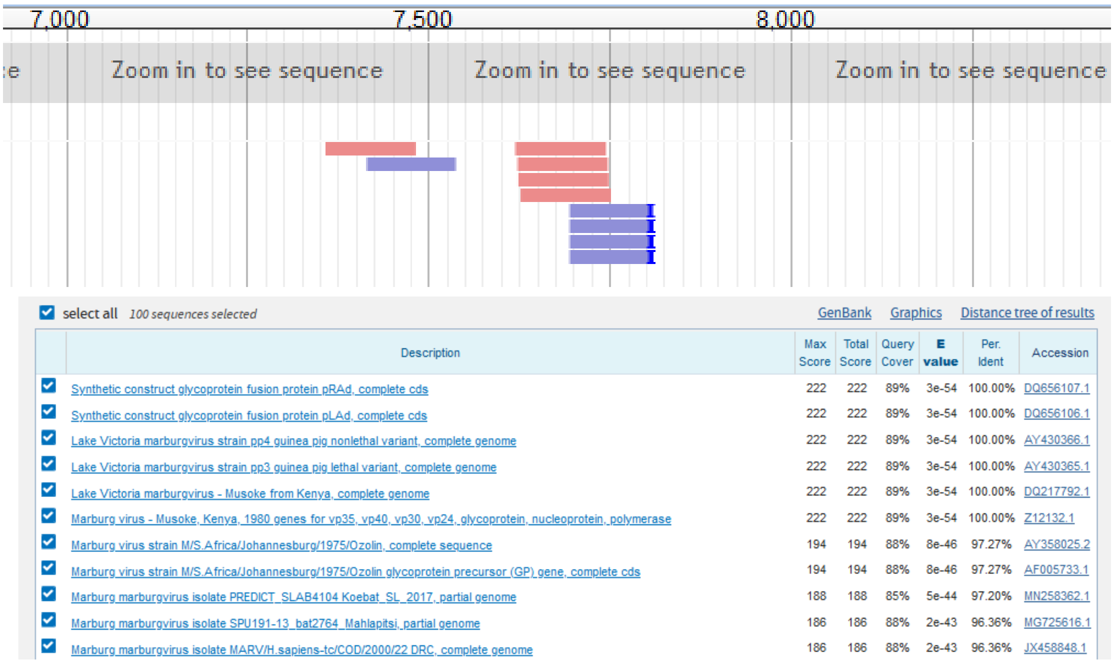
Database	nt See details
Query ID	lcl|Query_4887
Description	None
Molecule type	dna
Query Length	84
Other reports	❓

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

Fig.18b: No matches were obtained on the sequences 5’ to the S2m motif found in these reads.

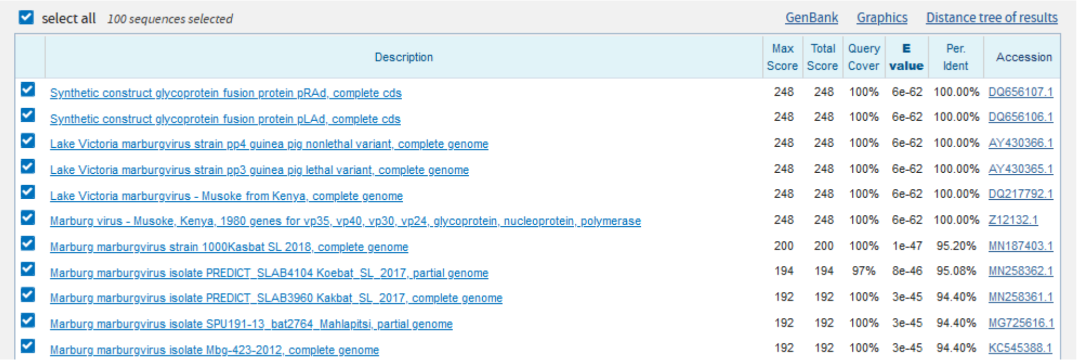
Presence of Type Strain Marburg Marburgvirus fragment in SRR9644024.

We obtained a total of 10 reads from Filoviridae, all of which matches 100% to a type strain Marburg Marburgvirus isolated in 1980 in Africa, of which no other isolates share the same sequence of nucleotides except for a synthetic construct for the Glycoprotein Fusion protein of Marburg Marburgvirus used originally for the vaccination of Guinea pigs in 2006[6].



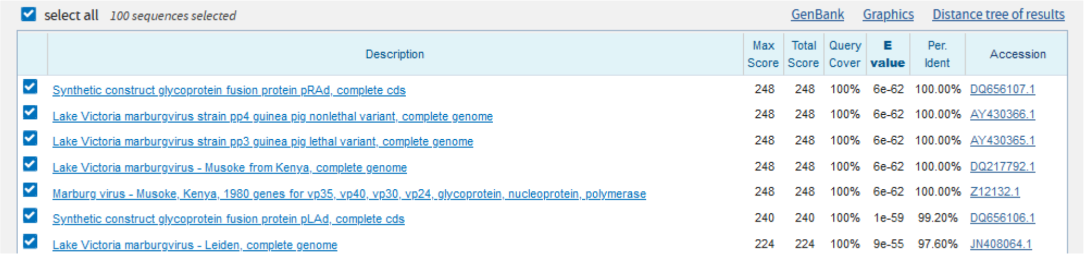
☑ select all 100 sequences selected

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Synthetic construct glycoprotein fusion protein pRAd, complete cds	222	222	89%	3e-54	100.00%	DQ656107.1
☑	Synthetic construct glycoprotein fusion protein pLAd, complete cds	222	222	89%	3e-54	100.00%	DQ656106.1
☑	Lake Victoria marburgvirus strain pp4 guinea pig nonlethal variant, complete genome	222	222	89%	3e-54	100.00%	AY430366.1
☑	Lake Victoria marburgvirus strain pp3 guinea pig lethal variant, complete genome	222	222	89%	3e-54	100.00%	AY430365.1
☑	Lake Victoria marburgvirus - Musoke from Kenya, complete genome	222	222	89%	3e-54	100.00%	DQ217792.1
☑	Marburg virus - Musoke, Kenya, 1980 genes for vp35, vp40, vp30, vp24, glycoprotein, nucleoprotein, polymerase	222	222	89%	3e-54	100.00%	Z12132.1
☑	Marburg virus strain M/S.Africa/Johannesburg/1975/Ozolin, complete sequence	194	194	88%	8e-46	97.27%	AY358025.2
☑	Marburg virus strain M/S.Africa/Johannesburg/1975/Ozolin glycoprotein precursor (GP) gene, complete cds	194	194	88%	8e-46	97.27%	AF005733.1
☑	Marburg marburgvirus isolate PREDICT_SLAB4104 Koebat_SL_2017, partial genome	188	188	85%	5e-44	97.20%	MN258362.1
☑	Marburg marburgvirus isolate SPU191-13_bat2764_Mahlapitsi, partial genome	186	186	88%	2e-43	96.36%	MG725616.1
☑	Marburg marburgvirus isolate MARV/H.sapiens-tc/COD/2000/22 DRC, complete genome	186	186	88%	2e-43	96.36%	JX458848.1



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

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Synthetic construct glycoprotein fusion protein pRAd, complete cds	248	248	100%	6e-62	100.00%	DQ656107.1
☑	Synthetic construct glycoprotein fusion protein pLAd, complete cds	248	248	100%	6e-62	100.00%	DQ656106.1
☑	Lake Victoria marburgvirus strain pp4 guinea pig nonlethal variant, complete genome	248	248	100%	6e-62	100.00%	AY430366.1
☑	Lake Victoria marburgvirus strain pp3 guinea pig lethal variant, complete genome	248	248	100%	6e-62	100.00%	AY430365.1
☑	Lake Victoria marburgvirus - Musoke from Kenya, complete genome	248	248	100%	6e-62	100.00%	DQ217792.1
☑	Marburg virus - Musoke, Kenya, 1980 genes for vp35, vp40, vp30, vp24, glycoprotein, nucleoprotein, polymerase	248	248	100%	6e-62	100.00%	Z12132.1
☑	Marburg marburgvirus strain 1000Kasbat SL 2018, complete genome	200	200	100%	1e-47	95.20%	MN187403.1
☑	Marburg marburgvirus isolate PREDICT_SLAB4104 Koebat_SL_2017, partial genome	194	194	97%	8e-46	95.08%	MN258362.1
☑	Marburg marburgvirus isolate PREDICT_SLAB3960 Kakbat_SL_2017, complete genome	192	192	100%	3e-45	94.40%	MN258361.1
☑	Marburg marburgvirus isolate SPU191-13_bat2764_Mahlapitsi, partial genome	192	192	100%	3e-45	94.40%	MG725616.1
☑	Marburg marburgvirus isolate Mbg-423-2012, complete genome	192	192	100%	3e-45	94.40%	KC545388.1

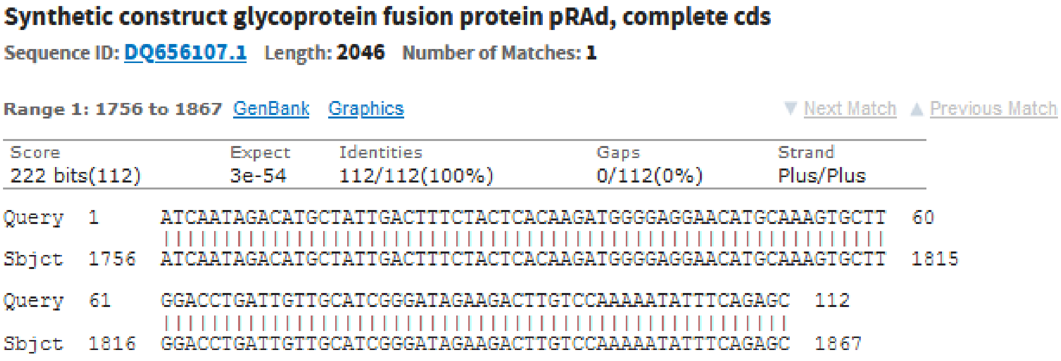


☑ select all *100 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Synthetic construct glycoprotein fusion protein pRAd, complete cds	248	248	100%	6e-62	100.00%	DQ656107.1
☑	Lake Victoria marburgvirus strain pp4 guinea pig nonlethal variant, complete genome	248	248	100%	6e-62	100.00%	AY430366.1
☑	Lake Victoria marburgvirus strain pp3 guinea pig lethal variant, complete genome	248	248	100%	6e-62	100.00%	AY430365.1
☑	Lake Victoria marburgvirus - Musoke from Kenya, complete genome	248	248	100%	6e-62	100.00%	DQ217792.1
☑	Marburg virus - Musoke, Kenya, 1980 genes for vp35, vp40, vp30, vp24, glycoprotein, nucleoprotein, polymerase	248	248	100%	6e-62	100.00%	Z12132.1
☑	Synthetic construct glycoprotein fusion protein pLAd, complete cds	240	240	100%	1e-59	99.20%	DQ656106.1
☑	Lake Victoria marburgvirus - Leiden, complete genome	224	224	100%	9e-55	97.60%	JN408064.1

Fig.19: Fragments of Marburg marburgvirus matching an 1980 type strain found in SRR9644024. As these reads does not have exact matches in any other field isolates of Marburg Marburgvirus, the most plausible explanation of such reads are the result of contamination from in-house vectors containing the fusion protein gene.

We also attempted to obtain a match to the last 14bp of the truncated reads at the extreme 3'-end of the alignment, However we could not find any meaningful matches on the NCBI database.



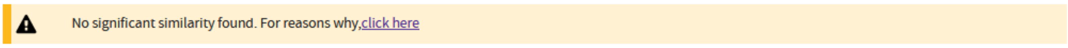
Synthetic construct glycoprotein fusion protein pRAd, complete cds

Sequence ID: DQ656107.1 Length: 2046 Number of Matches: 1

Range 1: 1756 to 1867 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
222 bits(112)	3e-54	112/112(100%)	0/112(0%)	Plus/Plus

```
Query  1     ATCAATAGACATGCTATTGACTTTCTACTCACAAGATGGGGAGGAACATGCAAAGTGCTT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1756  ATCAATAGACATGCTATTGACTTTCTACTCACAAGATGGGGAGGAACATGCAAAGTGCTT  1815

Query  61    GGACCTGATTGTTGCATCGGGATAGAAGACTTGTCCAAAAATATTTCAGAGC  112
             ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1816  GGACCTGATTGTTGCATCGGGATAGAAGACTTGTCCAAAAATATTTCAGAGC  1867
```



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

Fig.20: A perfect match to the Synthetic construct for the Glycoprotein of a Type strain Marburg Marburgvirus originally isolated in Africa, 1980. The extreme 3'-end sequence can not be found on the NCBI database.

No assembable sequences of other viruses exist for SRR9644024

We obtained 88 reads covering 12% pangenome from Reoviridae in SRR9644024 from segment 1, 2 and 3 of Rotavirus A. We did not obtain any other part of the 11-segmented viral genome.

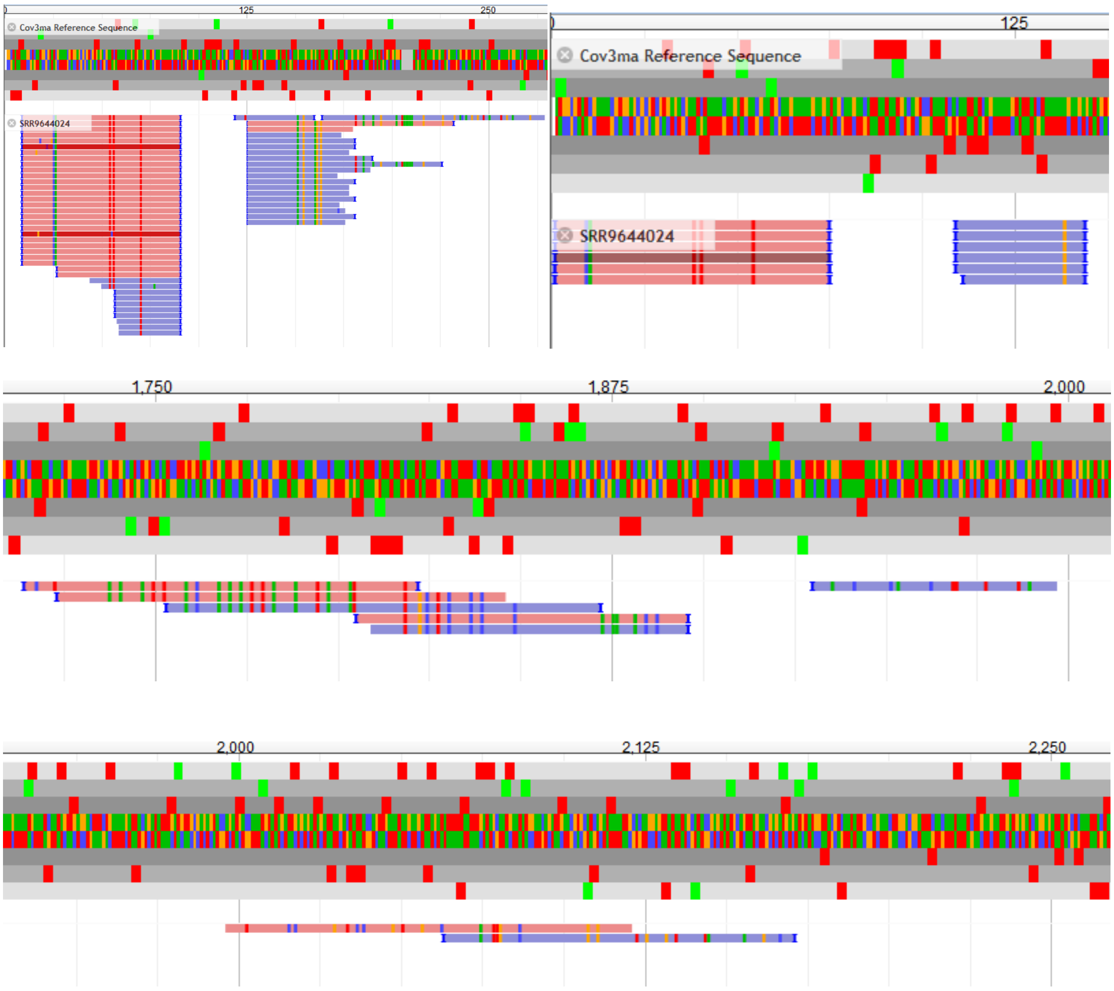


Fig.21a: reads from rotavirus A segment 2.

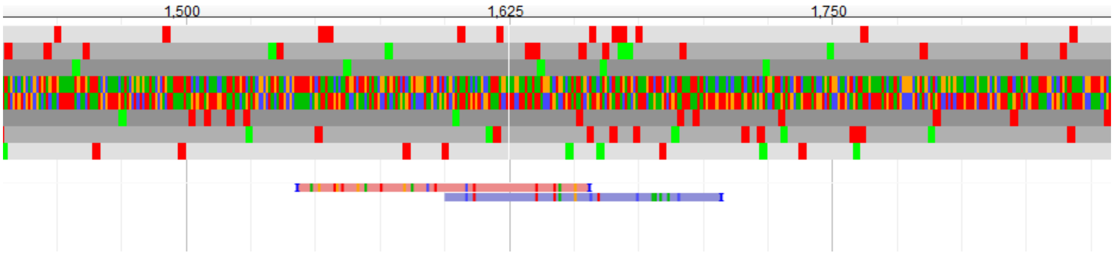


Fig.21b: reads from rotavirus A segment 1.

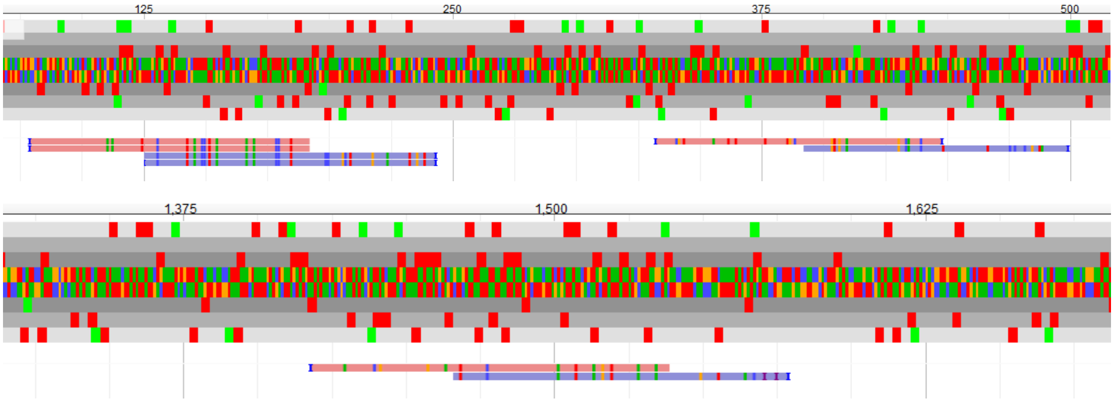
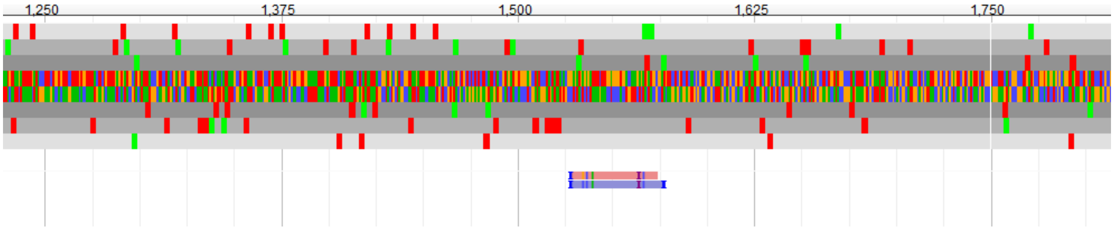
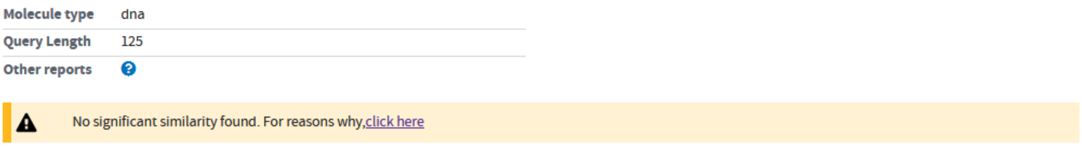


Fig.21c: reads from rotavirus A segment 3.

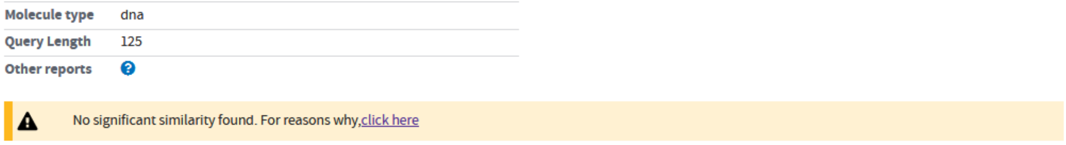
SERRATUS also claim alignment of 2 reads to Astroviruses, however these reads did not match anything when BLASTed.





Molecule type	dna
Query Length	125
Other reports	?

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



Molecule type	dna
Query Length	125
Other reports	?

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

Fig.21d: reads with claimed alignment to Astroviruses. However, these reads does not match anything when BLASTed.

MG-RAST analysis of SRR9644024 revealed significant levels of bacterial genomic DNA matching that found in other datasets submitted by the same group.

In order to elucidate the exact composition and nature of the mixed samples used to generate SRR9644024, We performed a MG-RAST analysis of SRR9644024. However, by using genomic DNA as the basis of search, we discovered that SRR9644024 contained 6.86% bacteria and up to 24.57% Caudovirales (Bacteriophages, which were often overrepresented in genomic DNA “total nucleic acids” as they were more highly annotated and have a denser coding region than the bacterial host which they integrate into.) which is opposed to the MG-RAST result of RaTG13, which contained only 3.94% bacteria and no evidence of Caudovirales.

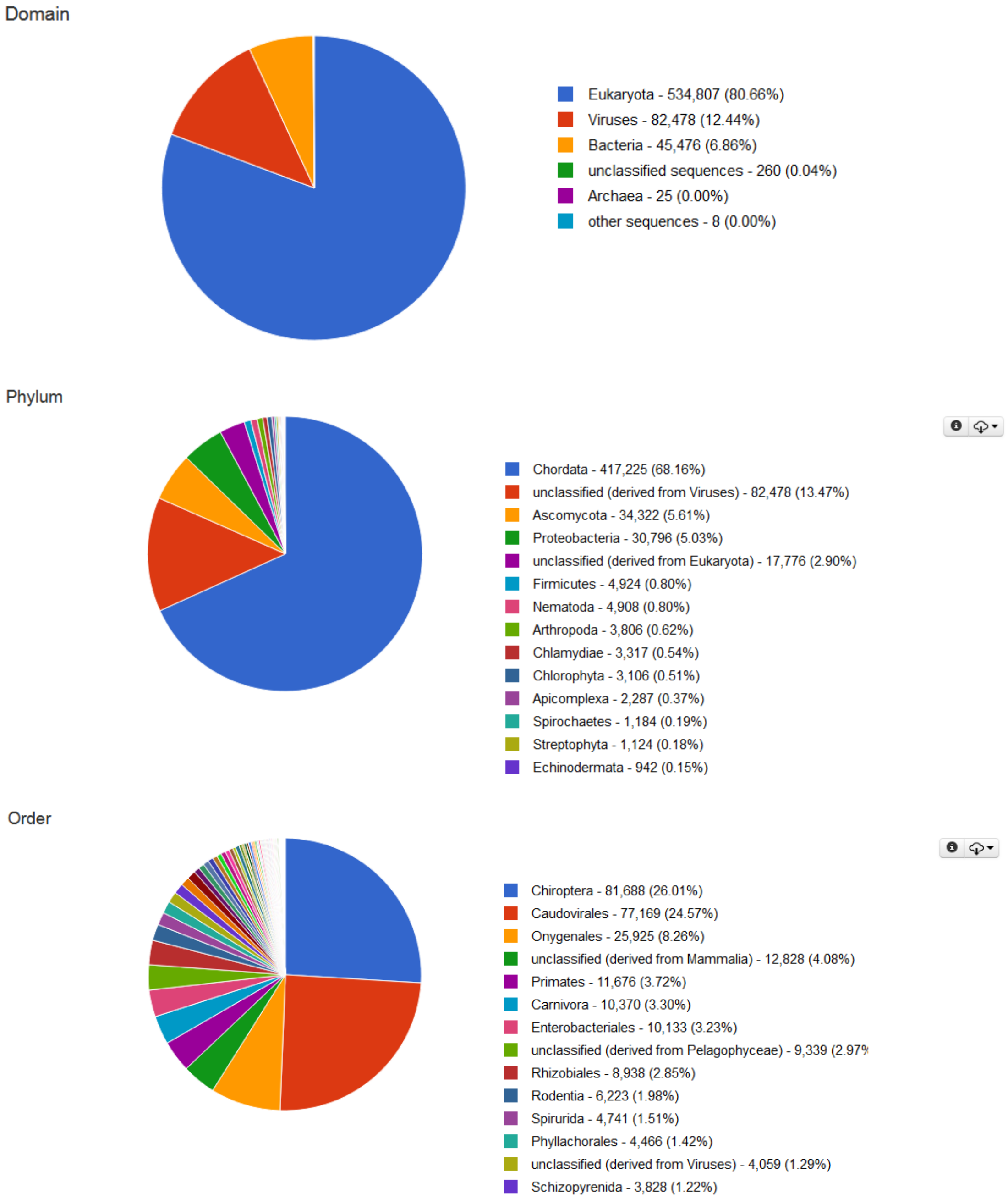


Fig.22a: MG-RAST analysis of SRR9644024 revealed bacterial reads of up to 6.86% in total genomic DNA. Caudovirales, representing prophages located within the bacterial genomes, contributed another 12.44% of the dataset.

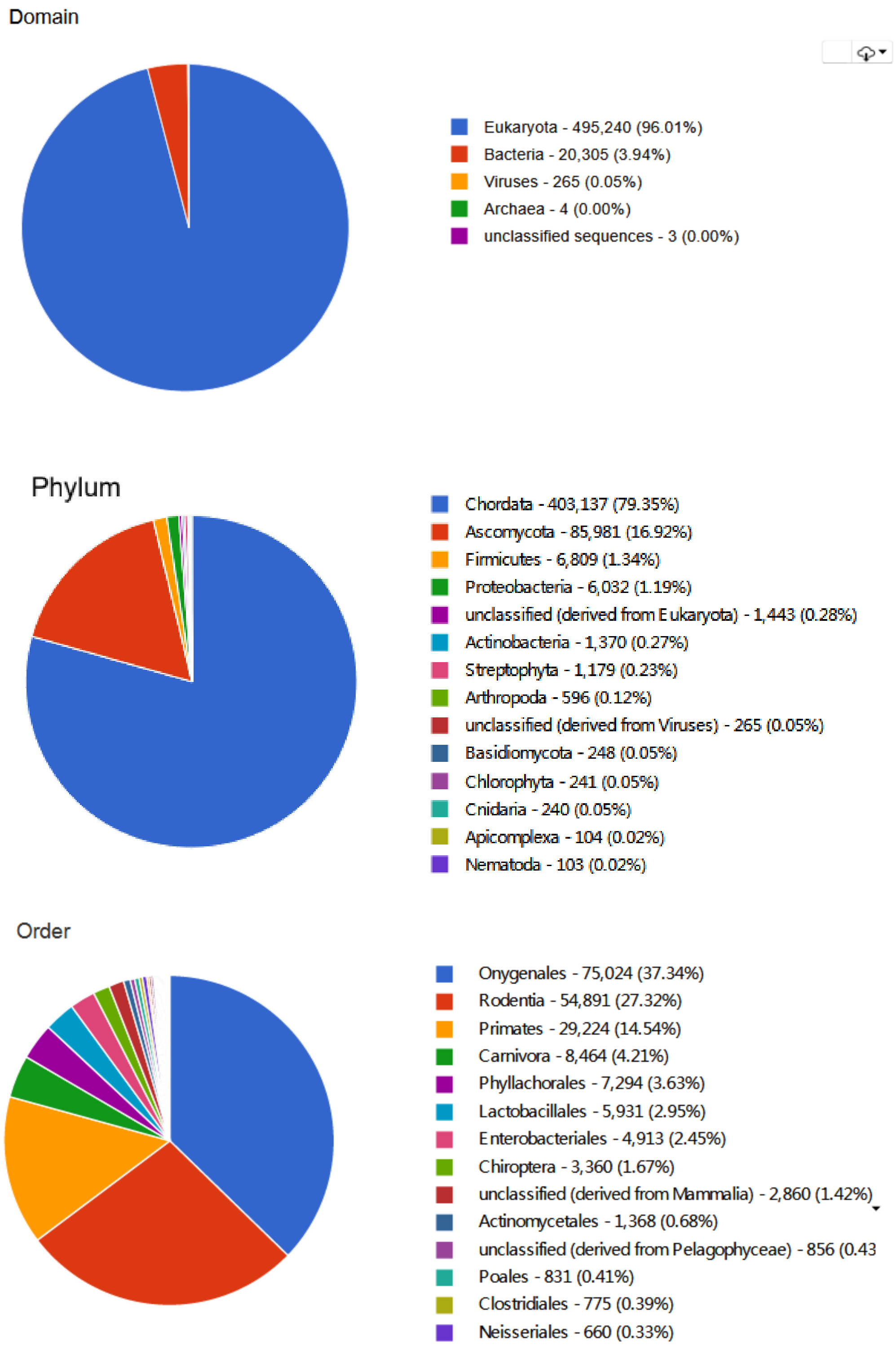
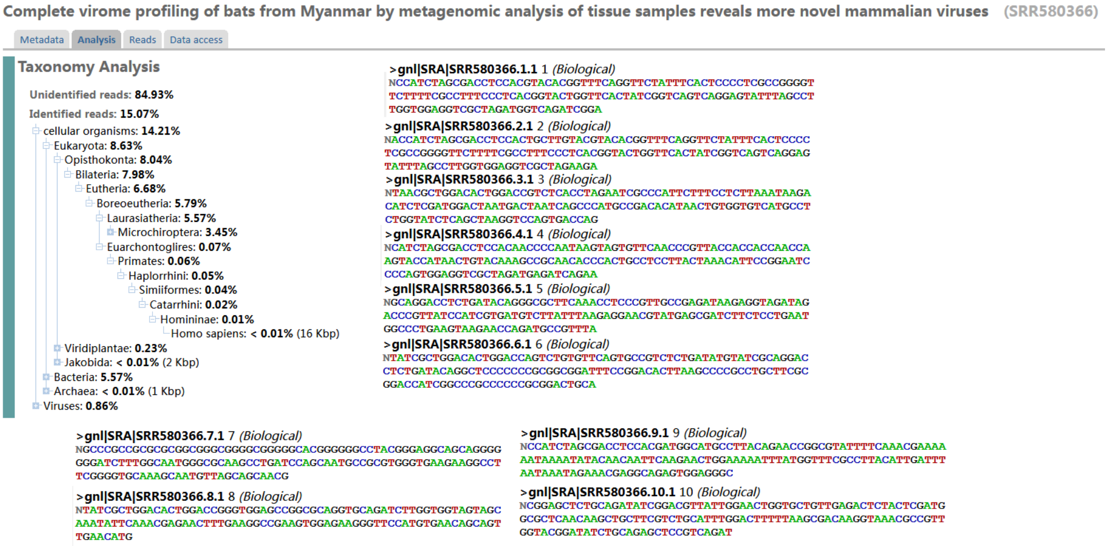


Fig.22b: MG-RAST analysis of RaTG13 revealed only 3.95% bacteria, and no sequences homologous to Caudovirales were found.

We also obtained 2 datasets prepared using "per previous methods" indicated in [4], However analysis of the datasets did not reveal any evidence of anomalies of either Telomere-like repeats or absence of bacteria in these datasets. [7]



Complete virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel mammalian viruses (SRR580366)

Taxonomy Analysis

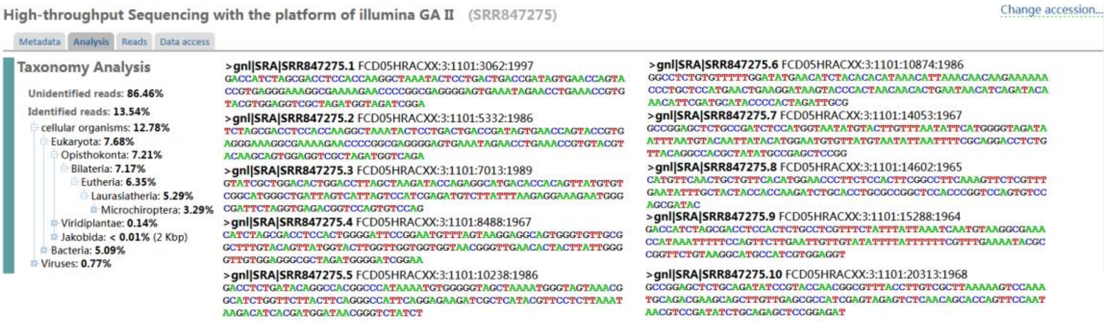
Unidentified reads: **84.93%**

Identified reads: **15.07%**

- cellular organisms: **14.21%**
  - Eukaryota: **8.63%**
    - Opisthokonta: **8.04%**
      - Bilateria: **7.98%**
        - Eutheria: **6.68%**
          - Boreoeutheria: **5.79%**
            - Laurasiatheria: **5.57%**
              - Microchiroptera: **3.45%**
            - Euarchontoglires: **0.07%**
              - Primates: **0.06%**
                - Haplorrhini: **0.05%**
                  - Simiiformes: **0.04%**
                    - Catarrhini: **0.02%**
                      - Homininae: **0.01%**
                        - Homo sapiens: < **0.01%** (16 Kbp)
    - Viridiplantae: **0.23%**
    - Jakobida: < **0.01%** (2 Kbp)
  - Bacteria: **5.57%**
  - Archaea: < **0.01%** (1 Kbp)
- Viruses: **0.86%**

```
>gnl|SRA|SRR580366.1.1 1 (Biological)
NCCATCTAGCGACCTCCACGTACACGGTTTCAGGTTCTATTTCACTCCCCTCGCCGGGGT
TCTTTTCGCCTTTCCCTCACGGTACTGGTTCACTATCGGTCAGTCAGGAGTATTTAGCCT
TGGTGGAGGTCGCTAGATGGTCAGATCGGA
>gnl|SRA|SRR580366.2.1 2 (Biological)
NACCATCTAGCGACCTCCACTGCTTGTACGTACACGGTTTCAGGTTCTATTTCACTCCCC
TCGCCGGGGTTCTTTTCGCCTTTCCCTCACGGTACTGGTTCACTATCGGTCAGTCAGGAG
TATTTAGCCTTGGTGGAGGTCGCTAGAAGA
>gnl|SRA|SRR580366.3.1 3 (Biological)
NTAACGCTGGACACTGGACCGTCTCACCTAGAATCGCCCATTCTTTCCTCTTAAATAAGA
CATCTCGATGGACTAATGACTAATCAGCCCATGCCGACACATAACTGTGGTGTCATGCCT
CTGGTATCTCAGCTAAGGTCCAGTGACCAG
>gnl|SRA|SRR580366.4.1 4 (Biological)
NCATCTAGCGACCTCCACAACCCCAATAAGTAGTGTTCAACCCGTTACCACCACCAACCA
AGTACCATAACTGTACAAAGCCGCAACACCCACTGCCTCCTTACTAAACATTCCGGAATC
CCCAGTGGAGGTCGCTAGATGAGATCAGAA
>gnl|SRA|SRR580366.5.1 5 (Biological)
NGCAGGACCTCTGATACAGGGCGCTTCAAACCTCCCGTTGCCGAGATAAGAGGTAGATAG
ACCCGTTATCCATCGTGATGTCTTATTTAAGAGGAACGTATGAGCGATCTTCTCCTGAAT
GGCCCTGAAGTAAGAACCAGATGCCGTTTA
>gnl|SRA|SRR580366.6.1 6 (Biological)
NTATCGCTGGACACTGGACCAGTCTGTGTTCAGTGCCGTCTCTGATATGTATCGCAGGAC
CTCTGATACAGGCTCCCCCCCGCGGCGGATTTCCGGACACTTAAGCCCCGCCTGCTTCGC
GGACCATCGGCCCGCCCCCCGCGGACTGCA
>gnl|SRA|SRR580366.7.1 7 (Biological)
NGCCCGCCGCGCGCGGCGGGCGGGGCGGGGGCACGGGGGGCCTACGGGAGGCAGCAGGGG
GGGATCTTTGGCAATGGGCGCAAGCCTGATCCAGCAATGCCGCGTGGGTGAAGAAGGCCT
TCGGGGTGCAAAGCAATGTTAGCAGCAACG
>gnl|SRA|SRR580366.8.1 8 (Biological)
NTATCGCTGGACACTGGACCGGGTGGAGCCGGCGCAGGTGCAGATCTTGGTGGTAGTAGC
AAATATTCAAACGAGAACTTTGAAGGCCGAAGTGGAGAAGGGTTCCATGTGAACAGCAGT
TGAACATG
>gnl|SRA|SRR580366.9.1 9 (Biological)
NCCATCTAGCGACCTCCACGATGGCATGCCTTACAGAACCGGCGTATTTTCAAACGAAAA
AATAAAATATACAACAATTCAAGAACTGGAAAAATTTATGGTTTCGCCTTACATTGATTT
AATAAATAGAAACGAGGCAGAGTGGAGGGC
>gnl|SRA|SRR580366.10.1 10 (Biological)
NCGGAGCTCTGCAGATATCGGACGTTATTGGAACTGGTGCTGTTGAGACTCTACTCGATG
GCGCTCAACAAGCTGCTTCGTCTGCATTTGGACTTTTTAAGCGACAAGGTAAACGCCGTT
GGTACGGATATCTGCAGAGCTCCGTCAGAT
```

Fig.24a: Analysis of SRR580366



High-throughput Sequencing with the platform of illumina GA II (SRR847275)

Taxonomy Analysis

Unidentified reads: **86.46%**

Identified reads: **13.54%**

- cellular organisms: **12.78%**
  - Eukaryota: **7.68%**
    - Opisthokonta: **7.21%**
      - Bilateria: **7.17%**
        - Eutheria: **6.35%**
          - Laurasiatheria: **5.29%**
            - Microchiroptera: **3.29%**
    - Viridiplantae: **0.14%**
    - Jakobida: < **0.01%** (2 Kbp)
  - Bacteria: **5.09%**
- Viruses: **0.77%**

```
>gnl|SRA|SRR847275.1 FCD05HRACXX:3:1101:3062:1997
GACCATCTAGCGACCTCCACCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTA
CCGTGAGGGAAAGGCGAAAAGAACCCCGGCGAGGGGAGTGAAATAGAACCTGAAACCGTG
TACGTGGAGGTCGCTAGATGGTAGATCGGA
>gnl|SRA|SRR847275.2 FCD05HRACXX:3:1101:5332:1986
TCTAGCGACCTCCACCAAGGCTAAATACTCCTGACTGACCGATAGTGAACCAGTACCGTG
AGGGAAAGGCGAAAAGAACCCCGGCGAGGGGAGTGAAATAGAACCTGAAACCGTGTACGT
ACAAGCAGTGGAGGTCGCTAGATGGTCAGA
>gnl|SRA|SRR847275.3 FCD05HRACXX:3:1101:7013:1989
GTATCGCTGGACACTGGACCTTAGCTAAGATACCAGAGGCATGACACCACAGTTATGTGT
CGGCATGGGCTGATTAGTCATTAGTCCATCGAGATGTCTTATTTAAGAGGAAAGAATGGG
CGATTCTAGGTGAGACGGTCCAGTGTCCAG
>gnl|SRA|SRR847275.4 FCD05HRACXX:3:1101:8488:1967
CATCTAGCGACCTCCACTGGGGATTCCGGAATGTTTAGTAAGGAGGCAGTGGGTGTTGCG
GCTTTGTACAGTTATGGTACTTGGTTGGTGGTGGTAACGGGTTGAACACTACTTATTGGG
GTTGTGGAGGGCGCTAGATGGGGATCGGAA
>gnl|SRA|SRR847275.5 FCD05HRACXX:3:1101:10238:1986
GACCTCTGATACAGGCCACGGCCCATAAAATGTGGGGGTAGCTAAAATGGGTAGTAAACG
GCATCTGGTTCTTACTTCAGGGCCATTCAGGAGAAGATCGCTCATACGTTCCTCTTAAAT
AAGACATCACGATGGATAACGGGTCTATCT
>gnl|SRA|SRR847275.6 FCD05HRACXX:3:1101:10874:1986
GGCCTCTGTGTTTTTGGATATGAACATCTACACACATAAACATTAAACAACAAGAAAAAA
CCCTGCTCCATGAACTGAAGGATAAGTACCCACTAACAACACTGAATAACATCAGATACA
AACATTCGATGCATACCCCACTAGATTGCG
>gnl|SRA|SRR847275.7 FCD05HRACXX:3:1101:14053:1967
GCCGGAGCTCTGCCGATCTCCATGGTAATATGTACTTGTTTAATATTCATGGGGTAGATA
ATTTAATGTACAATTATACATGGAATGTGTTATGTAATATTAATTTTCGCAGGACCTCTG
TTACAGGCCACGCTATATGCCGAGCTCCGG
>gnl|SRA|SRR847275.8 FCD05HRACXX:3:1101:14602:1965
CATGTTCAACTGCTGTTCACATGGAACCCTTCTCCACTTCGGCCTTCAAAGTTCTCGTTT
GAATATTTGCTACTACCACCAAGATCTGCACCTGCGCCGGCTCCACCCGGTCCAGTGTCC
AGCGATAC
>gnl|SRA|SRR847275.9 FCD05HRACXX:3:1101:15288:1964
GACCATCTAGCGACCTCCACTCTGCCTCGTTTCTATTTATTAAATCAATGTAAGGCGAAA
CCATAAATTTTTCCAGTTCTTGAATTGTTGTATATTTTATTTTTTCGTTTGAAAATACGC
CGGTTCTGTAAGGCATGCCATCGTGGAGGT
>gnl|SRA|SRR847275.10 FCD05HRACXX:3:1101:20313:1968
GCCGGAGCTCTGCAGATATCCGTACCAACGGCGTTTACCTTGTCGCTTAAAAAGTCCAAA
TGCAGACGAAGCAGCTTGTTGAGCGCCATCGAGTAGAGTCTCAACAGCACCAGTTCCAAT
AACGTCCGATATCTGCAGAGCTCCGGAGAT
```

Fig.24b: Analysis of SRR847275. Analysis of SRA059263 and the 2 associated SRA datasets does not reveal any evidence of anomalies found in the SRA dataset of RaTG13.

DISCUSSIONS

Telomere-like repeats as signature of sample Tampering by the mixing of PCR products from another source into the sample that is to be analyzed.

In order to elucidate the reason behind the Telomere-like repeats fond in SRR9644024, we compared it against the non-anomalous dataset, SRR9643845, from the same publisher. We did not find any significant peculiarities within the viral reads obtained from this dataset.

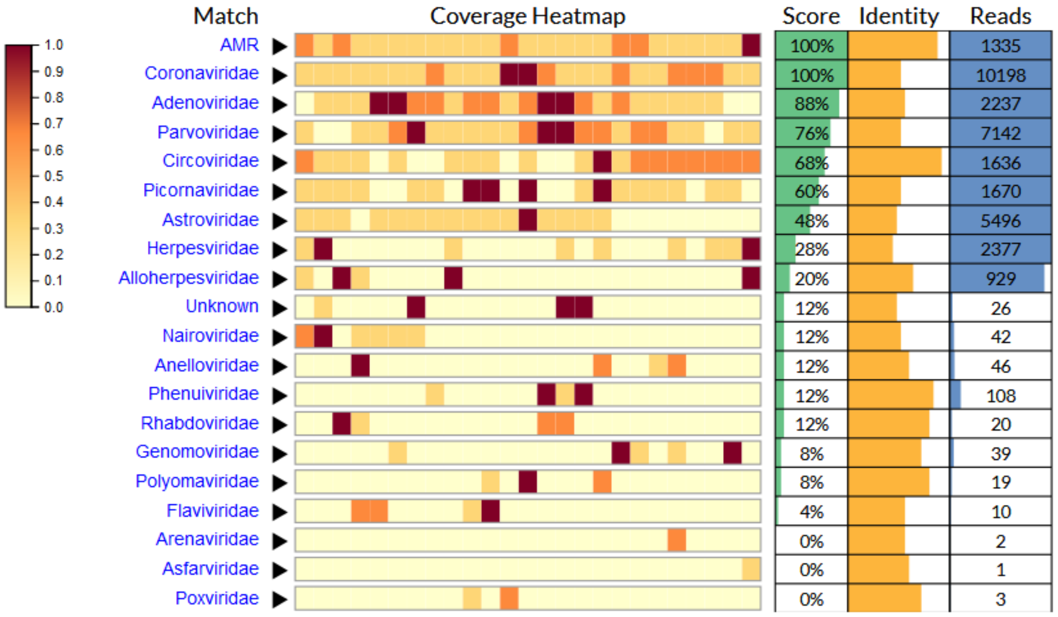
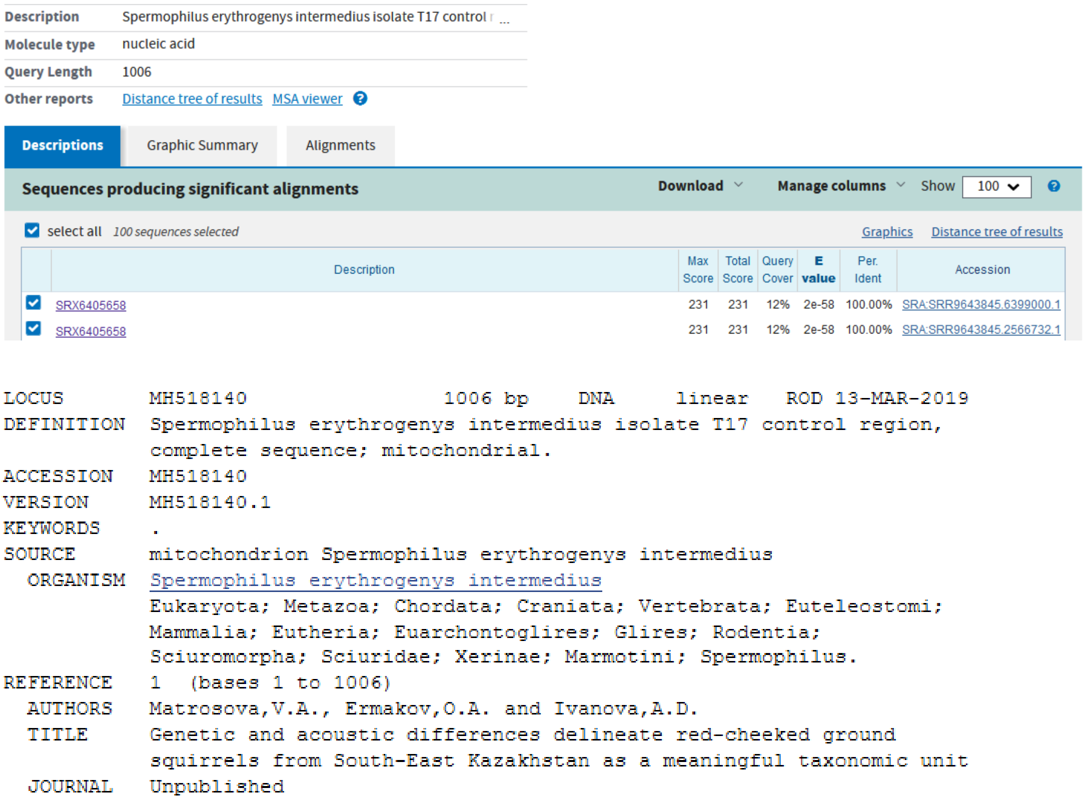


Fig.25: The SERRATUS analysis result of SRR9643845. No peculiarities of the viral reads were found.

The only peculiarity in SRR9643845 was the presence of reads from the Mitochondrial Control region of Spermophilus erythrogenys, which is a species of Marmots, alongside with Rattus Noverigicus and Homo Sapiens.



Description	Spermophilus erythrogenys intermedius isolate T17 control r ...
Molecule type	nucleic acid
Query Length	1006
Other reports	Distance tree of results MSA viewer

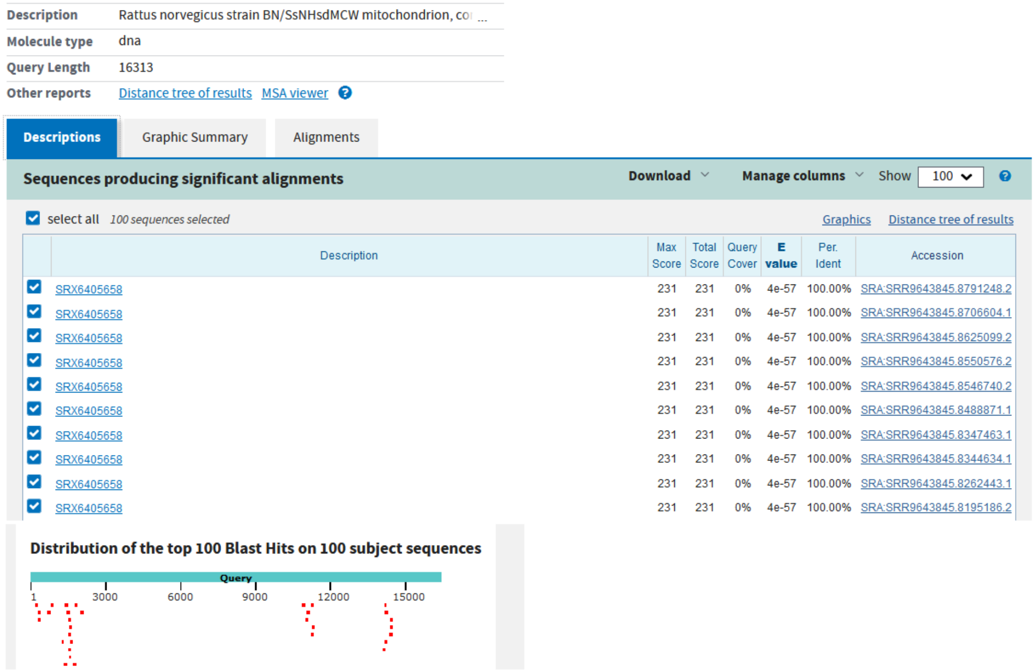
Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405658	231	231	12%	2e-58	100.00%	SRA:SRR9643845.6399000.1
☑	SRX6405658	231	231	12%	2e-58	100.00%	SRA:SRR9643845.2566732.1

```
LOCUS       MH518140                1006 bp    DNA     linear   ROD 13-MAR-2019
DEFINITION  Spermophilus erythrogenys intermedius isolate T17 control region,
            complete sequence; mitochondrial.
ACCESSION   MH518140
VERSION     MH518140.1
KEYWORDS    .
SOURCE      mitochondrion Spermophilus erythrogenys intermedius
  ORGANISM  Spermophilus erythrogenys intermedius
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciuromorpha; Sciuridae; Xerinae; Marmotini; Spermophilus.
REFERENCE   1  (bases 1 to 1006)
  AUTHORS   Matrosova,V.A., Ermakov,O.A. and Ivanova,A.D.
  TITLE     Genetic and acoustic differences delineate red-cheeked ground
            squirrels from South-East Kazakhstan as a meaningful taxonomic unit
  JOURNAL   Unpublished
```

Fig.26: Spermophilus erythrogenys Control region (D-loop) from SRR9643845. Spermophilus erythrogenys is a species of Marmots (Family: Marmotini).



Description	Rattus norvegicus strain BN/SsNHsdMCW mitochondrion, co ...
Molecule type	dna
Query Length	16313
Other reports	Distance tree of results MSA viewer

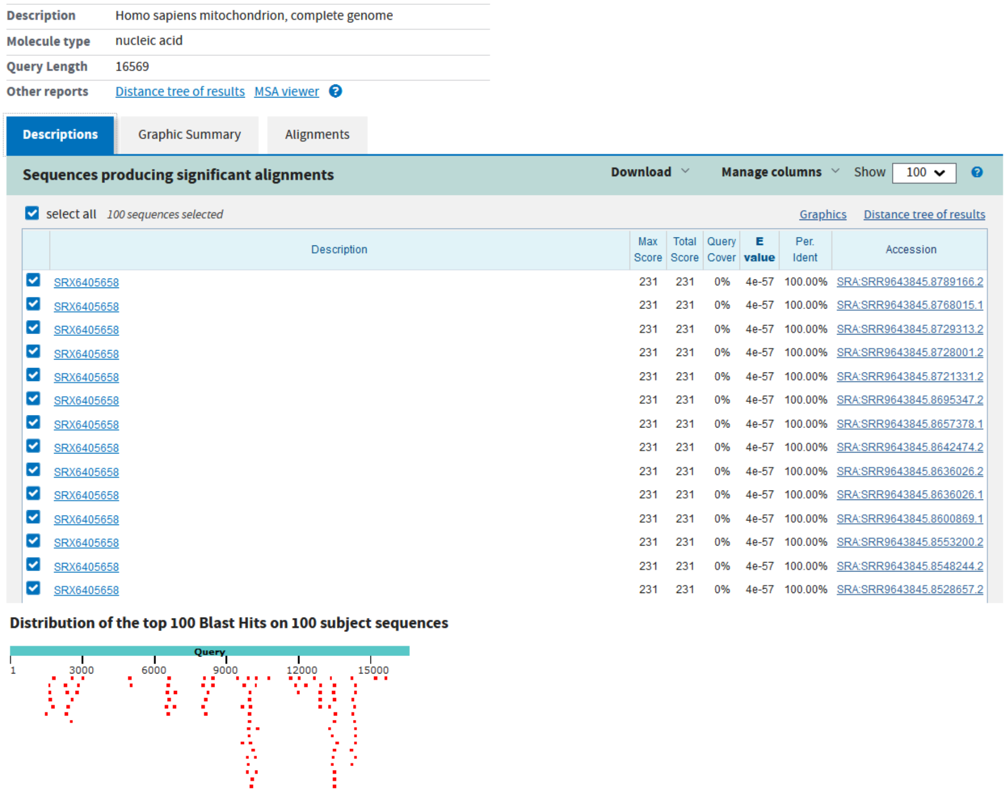
Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8791248.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8706604.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8625099.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8550576.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8546740.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8488871.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8347463.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8344634.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8262443.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8195186.2

Distribution of the top 100 Blast Hits on 100 subject sequences

Fig.27a: Rattus norvegicus Mitogenome recovered from SRR9643845.



Description	Homo sapiens mitochondrion, complete genome
Molecule type	nucleic acid
Query Length	16569
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
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8789166.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8768015.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8729313.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8728001.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8721331.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8695347.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8657378.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8642474.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8636026.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8636026.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8600869.1
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8553200.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8548244.2
☑	SRX6405658	231	231	0%	4e-57	100.00%	SRA:SRR9643845.8528657.2

Distribution of the top 100 Blast Hits on 100 subject sequences

Fig.27b: Homo Sapiens Mitogenome recovered from SRR9643845.

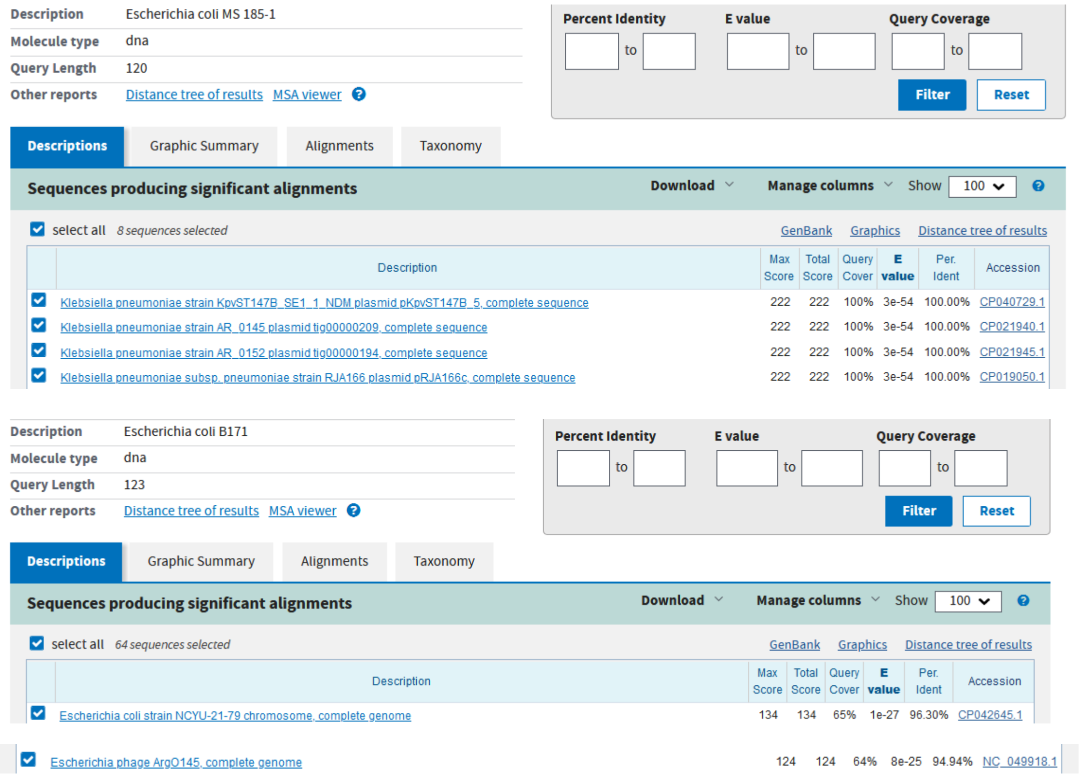
The only fundamental differences between SRR9644024 and SRR9643845 is that SRR9644024 contained a single amplicon sequence from Rabies Lyssavirus isolated from Mus Musculus, alongside with sequences that resembles primers that were stuck to DNA sequences of unknown origin. Using multiple SRA datasets from the same group as reference, the only plausible origins for the Telomere-like repeats in SRR9644024 is the numerous Type culture materials (Marburg

Marburgvirus) and the “rehosted” (from Mus Musculus) Amplicon of Rabies Lyssavirus.

In addition, Numerous reads resembling mispriming products by virus-specific primers on random DNA sequences can be found in SRR9644024, implying extended PCR amplification have been performed on multiple individual samples that were pooled into SRR9644024. Such extensive PCR manipulation resulted in the primer-independent amplification of trace repeat materials through the template sliding--reannealing mechanism, resulting in the formation of Telomere-like repeats in SRR9644024.

In addition, Type materials from cloning vectors like pRAd/DQ656107.1 are often extensively manipulated using PCR techniques, which can also lead to the amplification and accumulation of Telomere-like repeats in a sample containing such material.

Using MG-RAST results, we have confirmed the nature of SRR9644024 as a mixture of mostly specific PCR products from numerous sources—the bacterial reads were materials derived mostly from Prophages (Caudovirales) and Plasmids, while the Eukaryotic materials were mostly derived from Mitochondrial DNA.



Description	Escherichia coli MS 185-1
Molecule type	dna
Query Length	120
Other reports	Distance tree of results MSA viewer

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

Sequences producing significant alignments

☑ select all *8 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Klebsiella pneumoniae strain KpvST147B_SE1_1_NDM plasmid pKpvST147B_5, complete sequence	222	222	100%	3e-54	100.00%	CP040729.1
☑	Klebsiella pneumoniae strain AR_0145 plasmid tig00000209, complete sequence	222	222	100%	3e-54	100.00%	CP021940.1
☑	Klebsiella pneumoniae strain AR_0152 plasmid tig00000194, complete sequence	222	222	100%	3e-54	100.00%	CP021945.1
☑	Klebsiella pneumoniae subsp. pneumoniae strain RJA166 plasmid pRJA166c, complete sequence	222	222	100%	3e-54	100.00%	CP019050.1

Description	Escherichia coli B171
Molecule type	dna
Query Length	123
Other reports	Distance tree of results MSA viewer

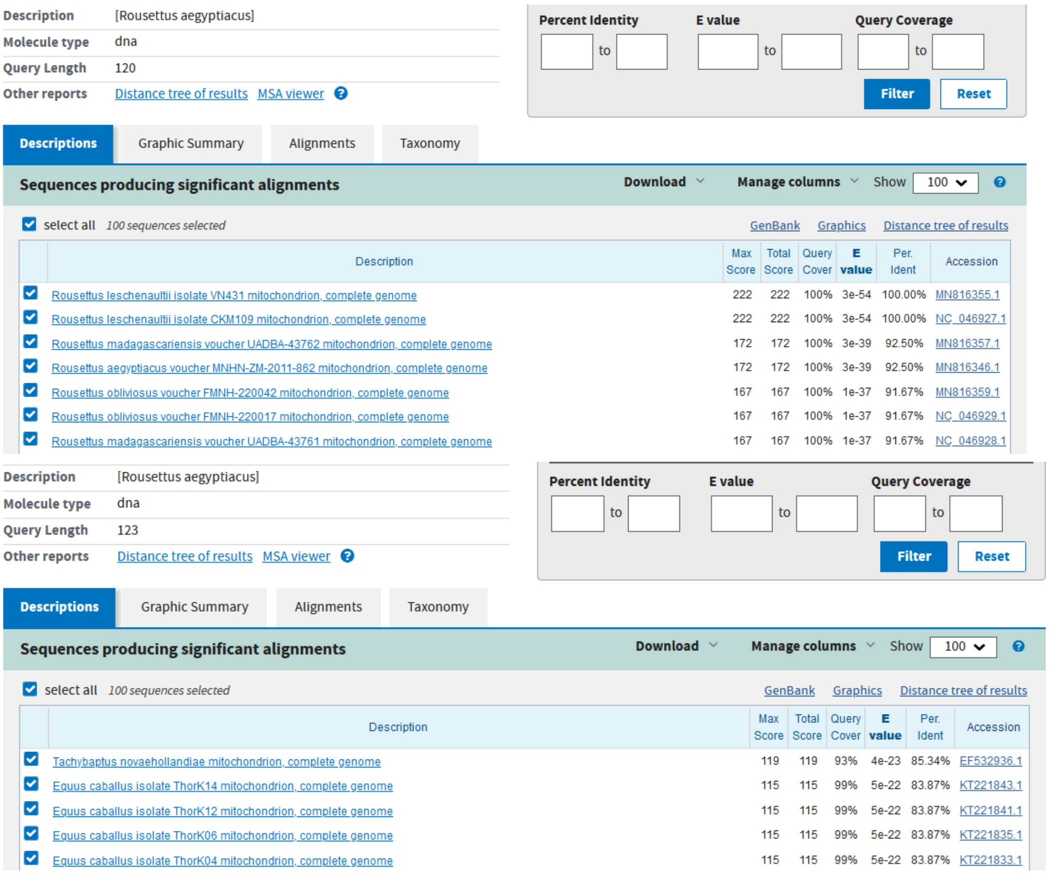
Percent Identity □ to □ E value □ to □ Query Coverage □ to □ Filter Reset

Sequences producing significant alignments

☑ select all *64 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Escherichia coli strain NCYU-21-79 chromosome, complete genome	134	134	65%	1e-27	96.30%	CP042645.1
☑	Escherichia phage ArgO145, complete genome	124	124	64%	8e-25	94.94%	NC_049918.1

Fig.28a: MG-RAST result of Bacteria in SRR9644024. These materials mostly matches to that of Prophages and Plasmids.



Description [Rousettus aegyptiacus]
Molecule type dna
Query Length 120
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
☑	Rousettus leschenaultii isolate VN431 mitochondrion, complete genome	222	222	100%	3e-54	100.00%	MN816355.1
☑	Rousettus leschenaultii isolate CKM109 mitochondrion, complete genome	222	222	100%	3e-54	100.00%	NC_046927.1
☑	Rousettus madagascariensis voucher UADBA-43762 mitochondrion, complete genome	172	172	100%	3e-39	92.50%	MN816357.1
☑	Rousettus aegyptiacus voucher MNHN-ZM-2011-862 mitochondrion, complete genome	172	172	100%	3e-39	92.50%	MN816346.1
☑	Rousettus obliviosus voucher FMNH-220042 mitochondrion, complete genome	167	167	100%	1e-37	91.67%	MN816359.1
☑	Rousettus obliviosus voucher FMNH-220017 mitochondrion, complete genome	167	167	100%	1e-37	91.67%	NC_046929.1
☑	Rousettus madagascariensis voucher UADBA-43761 mitochondrion, complete genome	167	167	100%	1e-37	91.67%	NC_046928.1

Description [Rousettus aegyptiacus]
Molecule type dna
Query Length 123
Other reports Distance tree of results MSA viewer

Percent Identity □ to □ E value □ to □ Query Coverage □ to □

Sequences producing significant alignments

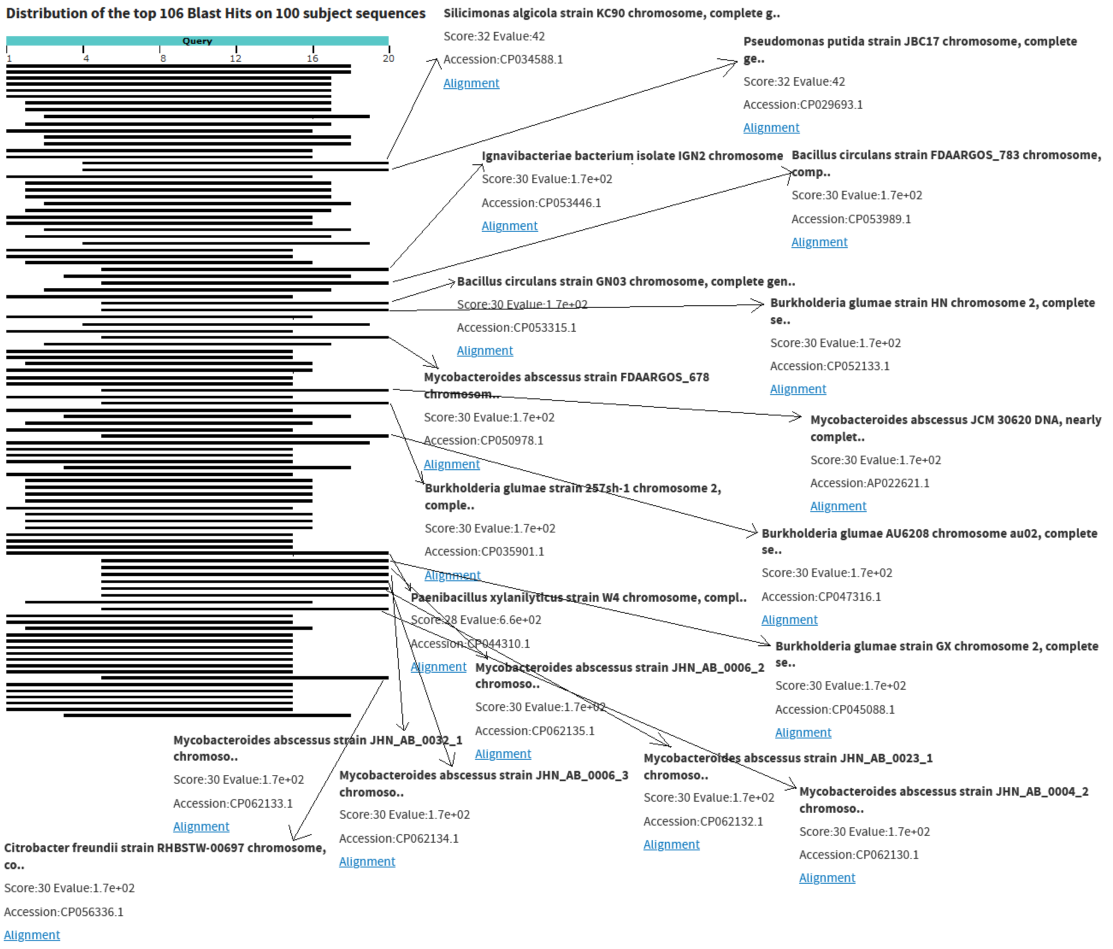
☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Tachybaptus novaehollandiae mitochondrion, complete genome	119	119	93%	4e-23	85.34%	EF532936.1
☑	Equus caballus isolate ThorK14 mitochondrion, complete genome	115	115	99%	5e-22	83.87%	KT221843.1
☑	Equus caballus isolate ThorK12 mitochondrion, complete genome	115	115	99%	5e-22	83.87%	KT221841.1
☑	Equus caballus isolate ThorK06 mitochondrion, complete genome	115	115	99%	5e-22	83.87%	KT221835.1
☑	Equus caballus isolate ThorK04 mitochondrion, complete genome	115	115	99%	5e-22	83.87%	KT221833.1

Fig.28b: MG-RAST result of Eukaryota in SRR9644024. These materials mostly match to Mitochondrial genomes.

We also identified a primer, 5’-GCCGGAGCTCTGCAGATATC-3’, used for the amplification of pooled total Nucleic acids in the preparation process of the library for SRR9644024, in the methods section from [5] at reference [7], [9] and [10] through cross-referencing. By performing a BLAST analysis, we discovered that this specific primer possessed significant bias against bacterial strains that are found to live on or within animals, which can cause significant depletion of bacteria if used, especially on animal samples which the microbiome is mostly composed of Epibiotic bacteria.

However, no evidence of the usage of such primer was found in the original paper for the sequencing of RaTG13. [11]



Distribution of the top 106 Blast Hits on 100 subject sequences

Silicimonas algicola strain KC90 chromosome, complete g..
Score:32 Evalue:42
Accession:CP034588.1
Alignment

Pseudomonas putida strain JBC17 chromosome, complete ge..
Score:32 Evalue:42
Accession:CP029693.1
Alignment

Ignavibacteriae bacterium isolate IGN2 chromosome
Score:30 Evalue:1.7e+02
Accession:CP053446.1
Alignment

Bacillus circulans strain FDAARGOS_783 chromosome, comp..
Score:30 Evalue:1.7e+02
Accession:CP053989.1
Alignment

Bacillus circulans strain GN03 chromosome, complete gen..
Score:30 Evalue:1.7e+02
Accession:CP053315.1
Alignment

Burkholderia glumae strain HN chromosome 2, complete se..
Score:30 Evalue:1.7e+02
Accession:CP052133.1
Alignment

Mycobacteroides abscessus strain FDAARGOS_678 chromosom..
Score:30 Evalue:1.7e+02
Accession:CP050978.1
Alignment

Mycobacteroides abscessus JCM 30620 DNA, nearly complet..
Score:30 Evalue:1.7e+02
Accession:AP022621.1
Alignment

Burkholderia glumae strain 257sh-1 chromosome 2, comple..
Score:30 Evalue:1.7e+02
Accession:CP035901.1
Alignment

Burkholderia glumae AU6208 chromosome au02, complete se..
Score:30 Evalue:1.7e+02
Accession:CP047316.1
Alignment

Paenibacillus xylanilyticus strain W4 chromosome, compl..
Score:28 Evalue:6.6e+02
Accession:CP044310.1
Alignment

Burkholderia glumae strain GX chromosome 2, complete se..
Score:30 Evalue:1.7e+02
Accession:CP045088.1
Alignment

Mycobacteroides abscessus strain JHN_AB_0006_2 chromoso..
Score:30 Evalue:1.7e+02
Accession:CP062135.1
Alignment

Mycobacteroides abscessus strain JHN_AB_0032_1 chromoso..
Score:30 Evalue:1.7e+02
Accession:CP062133.1
Alignment

Mycobacteroides abscessus strain JHN_AB_0006_3 chromoso..
Score:30 Evalue:1.7e+02
Accession:CP062134.1
Alignment

Mycobacteroides abscessus strain JHN_AB_0023_1 chromoso..
Score:30 Evalue:1.7e+02
Accession:CP062132.1
Alignment

Mycobacteroides abscessus strain JHN_AB_0004_2 chromoso..
Score:30 Evalue:1.7e+02
Accession:CP062130.1
Alignment

Citrobacter freundii strain RHBSTW-00697 chromosome, co..
Score:30 Evalue:1.7e+02
Accession:CP056336.1
Alignment

Fig.29: 3'-end alignment of Primer 5'-GCCGGAGCTCTGCAGATATC-3' to different strains of bacteria. Bacterial species that show 3'-end alignment all belongs to soil, environmental or pathogenic bacteria, which is not normally expected for samples of animal origin.

Revelation of manipulated material in the case of SRR9644024 is likely incidental

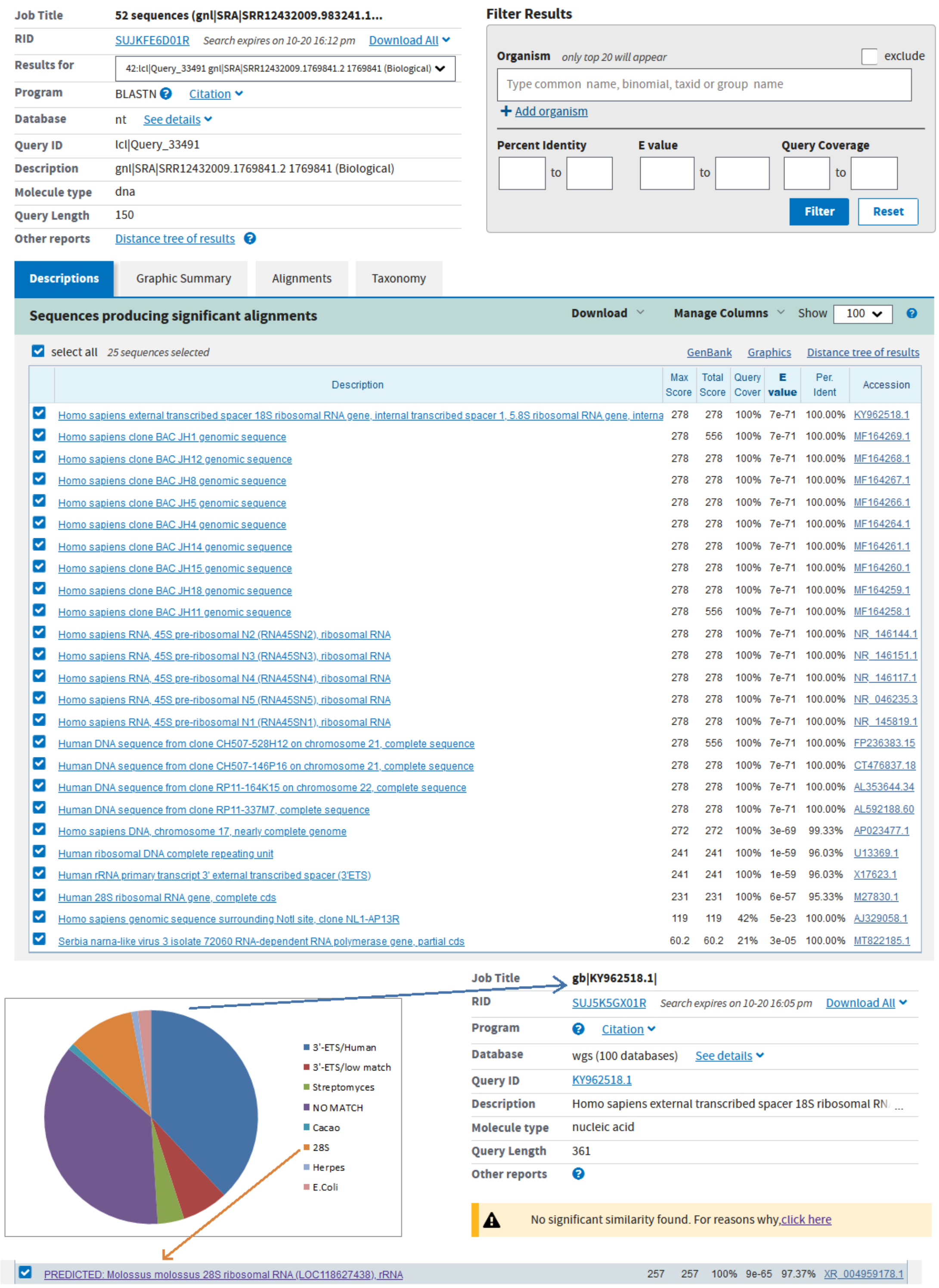
Although [4] and [5] was looking for polyomaviruses(PyVs) and are unlikely intentional in the manipulation of samples by themselves, their method section utilized archived samples from a large number of different studies (N>1000), which gives rise to a significant chance that material from yet unpublished studies, as well as internal practice materials for the manipulation and fabrication of viral metagenomic datasets, were incidentally included in the pooled sample of SRR9644024. In deed, the vast majority of RNA viral reads within SRR9644024 per NCBI analysis, belongs to the single amplicon of Rabies Lyssavirus with Mus Musculus DNA at the 3' end, suggesting that manipulated material comprised the majority (>90%) of the total nucleotides in the SRR9644024 pooled library, while the remainder were composed of Lungs, Intestines and Rectal tissues of different bats that may have not been fully degraded, leading to an unexpectedly high diversity of the mitochondrial reads within SRR9644024.

As neither Mus Musculus genomic DNA nor Marmotini Mitochondrial DNA were included in the list of sampled species in the supplementary table S1 of [4], the presence of the former at the 3' end of the Rabies Lyssavirus amplicon in SRR9644024 and the presence of the latter in SRR9643845 are indicative of materials from unpublished studies were being utilized in the pooling and sequencing of SRR9644024 and SRR9643845.

The discovery of obvious evidence of sample and metagenomic manipulation in SRR9644024, therefore, represents an incidental leakage of unpublished product of internal work-in-progress or proof-of-concept projects of PCR-based metagenomic manipulation through the incidental inclusion into a large pooled library that then get published in an unrelated study. However, such an incidental leakage nevertheless still provides valuable intel into the in-house protocols in the otherwise highly opaque and secretive institutions like the Military Academy of Sciences, allowing the nature of the anomalies in the sequencing datasets such as RaTG13 [1] to be analyzed and their origins deduced as the result of PCR-based metagenomic manipulation and fabrication.

Analysis of RmYN02

We also analyzed SRR12432009, the dataset for RmYN02 by individually retrieving 100 random reads from the dataset and then putting it through BLAST analysis. We discovered that nearly half of the dataset is composed of a single 3'-ETS sequence from Homo Sapiens, that does not have any matches in Chiroptera or Bats. Apart from data that can not be matched to anything on GenBank, SRR12432009 is composed of mostly parts of ribosomal RNA and contained about 6% bacterial sequences forming the rest of the identifiable reads within the dataset. We did not obtain significant matches to transcribed mRNA in SRR12432009.



Job Title	52 sequences (gnl|SRA|SRR12432009.983241.1...
RID	SUJKFE6D01R *Search expires on 10-20 16:12 pm* Download All
Results for	42:lcl|Query_33491 gnl|SRA|SRR12432009.1769841.2 1769841 (Biological)
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_33491
Description	gnl|SRA|SRR12432009.1769841.2 1769841 (Biological)
Molecule type	dna
Query Length	150
Other reports	Distance tree of results

Sequences producing significant alignments

☑ select all *25 sequences selected* GenBank Graphics Distance tree of results

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Homo sapiens external transcribed spacer 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, interna	278	278	100%	7e-71	100.00%	KY962518.1
☑	Homo sapiens clone BAC JH1 genomic sequence	278	556	100%	7e-71	100.00%	MF164269.1
☑	Homo sapiens clone BAC JH12 genomic sequence	278	278	100%	7e-71	100.00%	MF164268.1
☑	Homo sapiens clone BAC JH8 genomic sequence	278	278	100%	7e-71	100.00%	MF164267.1
☑	Homo sapiens clone BAC JH5 genomic sequence	278	278	100%	7e-71	100.00%	MF164266.1
☑	Homo sapiens clone BAC JH4 genomic sequence	278	278	100%	7e-71	100.00%	MF164264.1
☑	Homo sapiens clone BAC JH14 genomic sequence	278	278	100%	7e-71	100.00%	MF164261.1
☑	Homo sapiens clone BAC JH15 genomic sequence	278	278	100%	7e-71	100.00%	MF164260.1
☑	Homo sapiens clone BAC JH18 genomic sequence	278	278	100%	7e-71	100.00%	MF164259.1
☑	Homo sapiens clone BAC JH11 genomic sequence	278	556	100%	7e-71	100.00%	MF164258.1
☑	Homo sapiens RNA, 45S pre-ribosomal N2 (RNA45SN2), ribosomal RNA	278	278	100%	7e-71	100.00%	NR_146144.1
☑	Homo sapiens RNA, 45S pre-ribosomal N3 (RNA45SN3), ribosomal RNA	278	278	100%	7e-71	100.00%	NR_146151.1
☑	Homo sapiens RNA, 45S pre-ribosomal N4 (RNA45SN4), ribosomal RNA	278	278	100%	7e-71	100.00%	NR_146117.1
☑	Homo sapiens RNA, 45S pre-ribosomal N5 (RNA45SN5), ribosomal RNA	278	278	100%	7e-71	100.00%	NR_046235.3
☑	Homo sapiens RNA, 45S pre-ribosomal N1 (RNA45SN1), ribosomal RNA	278	278	100%	7e-71	100.00%	NR_145819.1
☑	Human DNA sequence from clone CH507-528H12 on chromosome 21, complete sequence	278	556	100%	7e-71	100.00%	FP236383.15
☑	Human DNA sequence from clone CH507-146P16 on chromosome 21, complete sequence	278	278	100%	7e-71	100.00%	CT476837.18
☑	Human DNA sequence from clone RP11-164K15 on chromosome 22, complete sequence	278	278	100%	7e-71	100.00%	AL353644.34
☑	Human DNA sequence from clone RP11-337M7, complete sequence	278	278	100%	7e-71	100.00%	AL592188.60
☑	Homo sapiens DNA, chromosome 17, nearly complete genome	272	272	100%	3e-69	99.33%	AP023477.1
☑	Human ribosomal DNA complete repeating unit	241	241	100%	1e-59	96.03%	U13369.1
☑	Human rRNA primary transcript 3' external transcribed spacer (3'ETS)	241	241	100%	1e-59	96.03%	X17623.1
☑	Human 28S ribosomal RNA gene, complete cds	231	231	100%	6e-57	95.33%	M27830.1
☑	Homo sapiens genomic sequence surrounding NotI site, clone NL1-AP13R	119	119	42%	5e-23	100.00%	AJ329058.1
☑	Serbia narna-like virus 3 isolate 72060 RNA-dependent RNA polymerase gene, partial cds	60.2	60.2	21%	3e-05	100.00%	MT822185.1

Job Title	gb|KY962518.1|
RID	SUJ5K5GX01R *Search expires on 10-20 16:05 pm* Download All
Program	Citation
Database	wgs (100 databases) See details
Query ID	KY962518.1
Description	Homo sapiens external transcribed spacer 18S ribosomal RN ...
Molecule type	nucleic acid
Query Length	361
Other reports	

No significant similarity found. For reasons why, click here

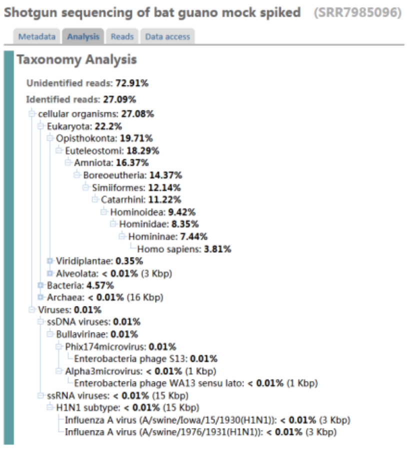
☑	PREDICTED: Molossus molossus 28S ribosomal RNA (LOC118627438), rRNA	257	257	100%	9e-65	97.37%	XR_004959178.1

Fig.30: Analysis of the RmYN02 dataset SRR12432009 using BLAST on 100 sequences randomly selected from the metagenomic sequencing dataset. We did not obtain any significant matches to bat mRNA and the reads were composed of mostly parts of the 45S ribosomal RNA cluster, with over 38% of all reads being exact matches to human DNA that does not have any significant matches in the current WGS dataset of Chiroptera(bats).

CONCLUSIONS

Through comparison between reference datasets and the only 2 datasets on NCBI that shares similar anomalies as the SRA data of RaTG13, We have deduced the origin of the Telomere-like sequences in RaTG13 as the result of mixing together PCR products from one virus (Rabies Lyssavirus isolated from Mus Musculus) into PCR products obtained from another (mostly degraded) sample, as Materials with obvious evidence of amplicons (SRR9644024) contained far greater concentration of such repeats than the degraded base material without being spiked with the amplicons (SRR975462), and other datasets obtained from the same method as SRR9644024 (SRR9643845, SRR580366 and SRR847275) failed to show evidence of anomalies.

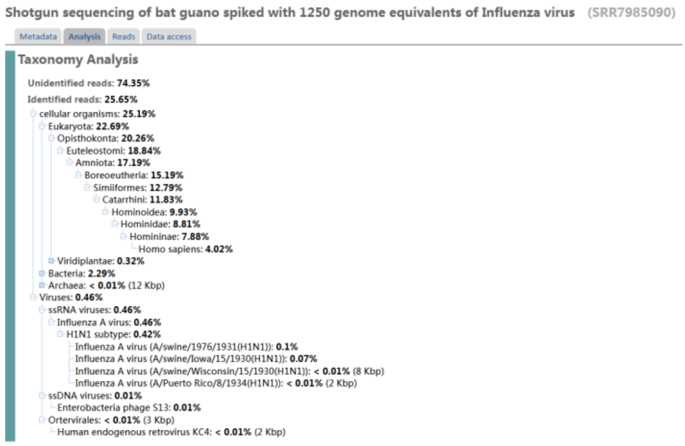
We also analyzed a metagenomic benchmark study which performed sequencing and analysis of different matrices Spiked with viral RNA, SRR7985096, SRR7985090 and SRR7985092. We discovered a trend of bacterial depletion as the amount of Spiked material is increased, which suggest that depletion of bacteria may also serve as a marker of sample manipulation, as manipulation of nucleic acid material invariably resulted in the degradation of original nucleic acids within the sample through various different processes.



Shotgun sequencing of bat guano mock spiked (SRR7985096)

Taxonomy Analysis

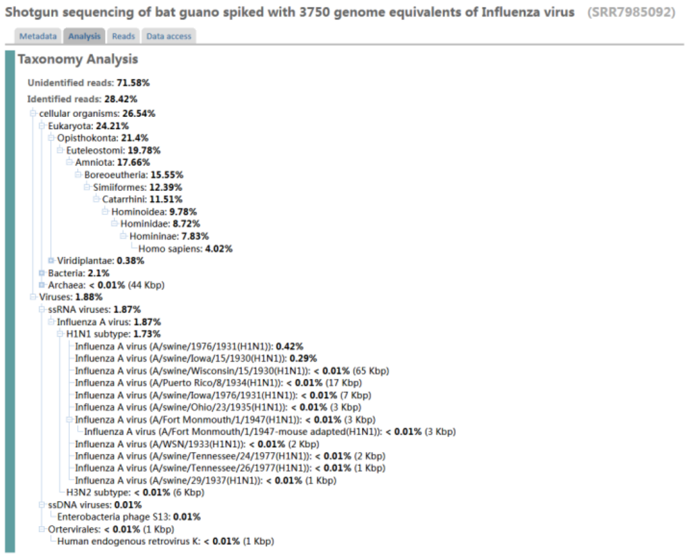
```
Unidentified reads: 72.91%
Identified reads: 27.09%
cellular organisms: 27.08%
  Eukaryota: 22.2%
    Opisthokonta: 19.71%
      Euteleostomi: 18.29%
        Amniota: 16.37%
          Boreoeutheria: 14.37%
            Simiiformes: 12.14%
              Catarrhini: 11.22%
                Hominoidea: 9.42%
                  Hominidae: 8.35%
                    Homininae: 7.44%
                      Homo sapiens: 3.81%
    Viridiplantae: 0.35%
    Alveolata: < 0.01% (3 Kbp)
  Bacteria: 4.57%
  Archaea: < 0.01% (16 Kbp)
Viruses: 0.01%
  ssDNA viruses: 0.01%
    Bullavirinae: 0.01%
      Phix174microvirus: 0.01%
        Enterobacteria phage S13: 0.01%
      Alpha3microvirus: < 0.01% (1 Kbp)
        Enterobacteria phage WA13 sensu lato: < 0.01% (1 Kbp)
  ssRNA viruses: < 0.01% (15 Kbp)
    H1N1 subtype: < 0.01% (15 Kbp)
      Influenza A virus (A/swine/Iowa/15/1930(H1N1)): < 0.01% (3 Kbp)
      Influenza A virus (A/swine/1976/1931(H1N1)): < 0.01% (3 Kbp)
```



Shotgun sequencing of bat guano spiked with 1250 genome equivalents of Influenza virus (SRR7985090)

Taxonomy Analysis

```
Unidentified reads: 74.35%
Identified reads: 25.65%
cellular organisms: 25.19%
  Eukaryota: 22.69%
    Opisthokonta: 20.26%
      Euteleostomi: 18.84%
        Amniota: 17.19%
          Boreoeutheria: 15.19%
            Simiiformes: 12.79%
              Catarrhini: 11.83%
                Hominoidea: 9.93%
                  Hominidae: 8.81%
                    Homininae: 7.88%
                      Homo sapiens: 4.02%
    Viridiplantae: 0.32%
  Bacteria: 2.29%
  Archaea: < 0.01% (12 Kbp)
Viruses: 0.46%
  ssRNA viruses: 0.46%
    Influenza A virus: 0.46%
      H1N1 subtype: 0.42%
        Influenza A virus (A/swine/1976/1931(H1N1)): 0.1%
        Influenza A virus (A/swine/Iowa/15/1930(H1N1)): 0.07%
        Influenza A virus (A/swine/Wisconsin/15/1930(H1N1)): < 0.01% (8 Kbp)
        Influenza A virus (A/Puerto Rico/8/1934(H1N1)): < 0.01% (2 Kbp)
  ssDNA viruses: 0.01%
    Enterobacteria phage S13: 0.01%
  Ortervirales: < 0.01% (3 Kbp)
    Human endogenous retrovirus KC4: < 0.01% (2 Kbp)
```



Shotgun sequencing of bat guano spiked with 3750 genome equivalents of Influenza virus (SRR7985092)

Taxonomy Analysis

```
Unidentified reads: 71.58%
Identified reads: 28.42%
cellular organisms: 26.54%
  Eukaryota: 24.21%
    Opisthokonta: 21.4%
      Euteleostomi: 19.78%
        Amniota: 17.66%
          Boreoeutheria: 15.55%
            Simiiformes: 12.39%
              Catarrhini: 11.51%
                Hominoidea: 9.78%
                  Hominidae: 8.72%
                    Homininae: 7.83%
                      Homo sapiens: 4.02%
    Viridiplantae: 0.38%
  Bacteria: 2.1%
  Archaea: < 0.01% (44 Kbp)
Viruses: 1.88%
  ssRNA viruses: 1.87%
    Influenza A virus: 1.87%
      H1N1 subtype: 1.73%
        Influenza A virus (A/swine/1976/1931(H1N1)): 0.42%
        Influenza A virus (A/swine/Iowa/15/1930(H1N1)): 0.29%
        Influenza A virus (A/swine/Wisconsin/15/1930(H1N1)): < 0.01% (65 Kbp)
        Influenza A virus (A/Puerto Rico/8/1934(H1N1)): < 0.01% (17 Kbp)
        Influenza A virus (A/swine/Iowa/1976/1931(H1N1)): < 0.01% (7 Kbp)
        Influenza A virus (A/swine/Ohio/23/1935(H1N1)): < 0.01% (3 Kbp)
        Influenza A virus (A/Fort Monmouth/1/1947(H1N1)): < 0.01% (3 Kbp)
          Influenza A virus (A/Fort Monmouth/1/1947-mouse adapted(H1N1)): < 0.01% (3 Kbp)
        Influenza A virus (A/WSN/1933(H1N1)): < 0.01% (2 Kbp)
        Influenza A virus (A/swine/Tennessee/24/1977(H1N1)): < 0.01% (2 Kbp)
        Influenza A virus (A/swine/Tennessee/26/1977(H1N1)): < 0.01% (1 Kbp)
        Influenza A virus (A/swine/29/1937(H1N1)): < 0.01% (1 Kbp)
      H3N2 subtype: < 0.01% (6 Kbp)
  ssDNA viruses: 0.01%
    Enterobacteria phage S13: 0.01%
  Ortervirales: < 0.01% (1 Kbp)
    Human endogenous retrovirus K: < 0.01% (1 Kbp)
```

Fig.31: Analysis of SRR7985096, SRR7985090 and SRR7985092. A trend of decreasing bacterial reads were observed when comparing the Mock Spiked (without viral RNA) material and material

Spiked with an increasing amount of viral RNA.

When the only 3 datasets on NCBI with the observed anomalies were compared against each other, a pipeline of metagenomic fabrication, involving the “rehosting” of viral reads from one sample to another through the Mixing in of PCR amplicons of the virus into a heavily degraded sample “matrix”, is clearly revealed: By adding a single amplicon of Rabies Lyssavirus from Mus Musculus into a mixture of degraded tissue samples similar to SRR9643845, A dataset similar to SRR9644024 is generated. By adding multiple amplicons from a plethora of different Coronaviruses into a degraded fecal sample similar to SRR975462, a dataset similar to the mNGS dataset of RaTG13 is generated.

Through comparative analysis of multiple datasets, we have also discovered the signature of such manipulation—the depletion of bacterial reads were the result of extensive sample manipulation destroying the original RNA within the matrix sample, while the enrichment of Telomere-like repeats is the result of spiking with material prepared using extensive, high cycle time PCR methods, especially those that are used for the manipulation of nucleotide sequences In Vitro.

We therefore urge all current studies that uses RaTG13 as the basis of argument on the origin of SARS-CoV-2 to be immediately revised and corrected.

REFERENCES

[1] Rahalkar, M.; Bahulikar, R. The Anomalous Nature of the Fecal Swab Data, Receptor Binding Domain and Other Questions in RaTG13 Genome . Preprints 2020, 2020080205 (doi: 10.20944/preprints202008.0205.v3).

[2] Daoyu Zhang. (2020, August 1). Anomalies in BatCoV/RaTG13 sequencing and provenance. Zenodo. http://doi.org/10.5281/zenodo.4064067

[3] Singla, M.; Ahmad, S.; Gupta, C.; Sethi, T. De-novo Assembly of RaTG13 Genome Reveals Inconsistencies Further Obscuring SARS-CoV-2 Origins. *Preprints* **2020**, 2020080595 (doi: 10.20944/preprints202008.0595.v1).

[4] Zhizhou Tan1¶,Gabriel Gonzalez2¶, Jinliang Sheng3, Jianmin Wu4, Fuqiang Zhang5, Lin Xu1, 6Peisheng Zhang1, 3, Aiwei Zhu1, Yonggang Qu3,Changchun Tu1,6, Michael J. Carr7#, Biao He1,6 Extensive genetic diversity of polyomaviruses in sympatric bat communities: host-switchingversusco-evolution J. Virol. doi:10.1128/JVI.02101-19

[5] Extensive genetic diversity of bat-borne polyomaviruses reveals inter-family host-switching events

Zhizhou Tan, Gabriel Gonzalez, Jinliang Sheng, Jianmin Wu, Fuqiang Zhang, Lin Xu, Peisheng Zhang, Aiwei Zhu, Yonggang Qu, Changchun Tu, Michael J. Carr, Biao He

bioRxiv 627158; doi: https://doi.org/10.1101/627158

[6] Wang D, Hevey M, Juompan LY, Trubey CM, Raja NU, Deitz SB, Woraratanadharm J, Luo M, Yu H, Swain BM, Moore KM, Dong JY. Complex adenovirus-vectored vaccine protects guinea pigs from three strains of Marburg virus challenges. Virology. 2006 Sep 30;353(2):324-32. doi: 10.1016/j.virol.2006.05.033. Epub 2006 Jul 3. PMID: 16820184.

[7] He B, Li Z, Yang F, Zheng J, Feng Y, Guo H, Li Y, Wang Y, Su N, Zhang F, Fan Q, Tu C. Virome profiling of bats from Myanmar by metagenomic analysis of tissue samples reveals more novel Mammalian viruses. PLoS One. 2013 Apr 22;8(4):e61950. doi: 10.1371/journal.pone.0061950.

Erratum in: PLoS One. 2013;8(6). doi:10.1371/annotation/68f77773-a2a0-4bfe-b5e6-950dc30b79f9. PMID: 23630620; PMCID: PMC3632529.

[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] Donaldson EF, Haskew AN, Gates JE, Huynh J, Moore CJ, Frieman MB. Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J Virol*. 2010;84(24):13004-13018. doi:10.1128/JVI.01255-10

[10] Allander T, Tammi MT, Eriksson M, Bjerkner A, Tiveljung-Lindell A, Andersson B. Cloning of a human parvovirus by molecular screening of respiratory tract samples [published correction appears in Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15712]. *Proc Natl Acad Sci U S A*. 2005;102(36):12891-12896. doi:10.1073/pnas.0504666102

[11] Zhou, P., Yang, X., Wang, X. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7