doi:10.20944/preprints202008.0205.v1

Short Note

The Abnormal Nature of the Fecal Swab Sample used for NGS Analysis of RaTG13 Genome Sequence Imposes a Question on the Correctness of the RaTG13 Sequence

Monali C. Rahalkar[1]* and Rahul A. Bahulikar[2]

[1]C2, Bioenergy group, MACS Agharkar Research Institute, G.G. Agarkar Road, Pune 411004, Maharashtra, India

[2]BAIF Development Research Foundation, Central Research Station, Urulikanchan, Pune 412202

*Corresponding author: monalirahalkar@aripune.org

Abstract:

RaTG13 is the next relative of SARS-CoV-2 derived from bat feces. The Illumina based NGS sequence of RaTG13 MN996532.1 was deposited on 27th Jan 2020 and the raw data, a little later on 13th Feb 2020 https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn]. The fecal swab sample shows abnormally high reads from eukaryotes which includes not only bats but other animals, as per the NCBI site. Also, comparison of the fecal swab to other bat fecal swabs deposited by the same group on the same date indicates that the fecal swab from which RaTG13 sequence was derived looked abnormal. The proportion of bacteria in this RNA Seq project was only 0.7% in contrast to 70-90% abundance in other fecal swabs from bats. Also, the amplicon sequencing done on the same sample showed large number of gaps and inconsistencies. This poses a question on the authenticity of the RaTG13 sequence also.

Keywords: RaTG13; SARS-COV-2; Illumina sequencing; amplicon sequencing; NGS; fecal swab

Covid-19 has been a devastating pandemic affecting more than nineteen million people in more than 200 countries and killing three quarter million people till now. SARS-CoV2, the virus responsible for the disease is most similar to RaTG13 (a bat derived virus) on the genomic level. RaTG13 has been known as the sister virus of SARS-CoV-2 as its shows 96.2% overall genomic similarity to CoV-2 genome (Zhou et al., 2020). RaTG13 has been widely used for various comparative experiments with that of SARS-CoV-2. This includes the capacity of its spike to bind to human ACE-2, its infective capacity, etc. RaTG13 genome is also used for calculations of the common ancestor and also for further calculations before how long RaTG13 and SARS-CoV-2 got separated, etc.

RaTG13 is described as the virus (not a real virus, but available as a sequence) from the RNA of a bat fecal swab collected in July 2013, from Tongguan mines in Yunnan. The old name of RaTG13 virus is CoV4991 (Ge et al., 2016). However, the sample appears to be over or not available to the scientific community as per a recent news investigation (2020). One main condition for using RaTG13 for all future experiments is that the sequence of this virus should be accurate and based on a good raw data.

RaTG13 never seemed to have existed before SARS-COV-2 was described, as the genome sequence was not available on NCBI before (Zhou et al., 2020) .The Illumina based NGS sequence of RaTG13 **MN996532.1** was deposited on 27th Jan 2020 and the raw data, a little later on 13th Feb 2020 https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn].

The earlier name of RaTG13 is CoV/4991. A 370 base RdRp fragment (KP378696.1) of CoV/4991 and showed highest similarity to SARS-CoV-2 RdRp fragment with only 3-5 bases different (NCBI blast analysis). Also, 4991 or RaTG13 has a great significance as it was recovered from the same site where a COVID-19 like disease occurred (2020, Rahalkar and Bahulikar, 2020). CoV 4991 is also the first and only SARS-like CoV associated with human pneumonia cases, before SARS-COV-2 (Rahalkar and Bahulikar, 2020).

Problems seen in the RAW DATA of RaTG13: Illumina sequence SRX7724752

Here are the basic discrepancies encountered after the analysis of the Illumina raw data https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn]:

1. The genome of RaTG13 is derived from a fecal or anal swab (MN996532.1). However in the Illumina sequencing description, SRX7724752, the sample is described to be of a BAL fluid (broncho alveolar lavage).

2. The total raw data is 3.3 Gb. After the Krona analysis it is seen that ~30% reads are unidentified (no matches) and only ~ 70% reads are identified. Out of the 70%, a vast majority i.e. 68% was contributed by eukaryotes (fig. 1). This is highly unusual as it is a fecal swab and the analysis of other bat fecal or anal swabs cannot show such high proportion of eukaryotic RNA.

3. Within the 68% eukaryote sequences, the bat sequences are about 36-40% (Fig 1a.), and rest of the 30% sequences are contributed by squirrels, flying foxes, foxes, and other types of animals (Fig.1 b). First of all, why would such high proportion of eukaryotic sequences appear in the RNA when it's a fecal swab? From where do these animal sequences come when it is supposed to be a *Rhinophus affinis* swab? Also, even though the *Rhinophus affinis* sequence may not be present in the database, why are they similar to so many bat sequences? Some of these bats are found only in Mexico or USA (Zhang, 2020).

4. The RNA Seq data shows extremely less abundance of bacteria, only 0.65%. This is far too less in comparison to other fecal or anal swab of bats, which show a very high proportion of bacterial sequences ~76-90% (Fig.2 and.3). SRA data of six other fecal swabs submitted by

the same group were used for comparison (data not shown). Bacteria are the highest constituents of a fecal sample.

5. The coronavirus sequence (RaTG13) contributes to only ~0.003% of the total sequence reads. These raw reads were used to build an almost complete assembly, though the overall coverage is very less ~8X. Though there were less overlaps in some regions there are only 2-3 gaps. The Wuhan Institute of Virology has recently described methods like probe-capture for getting the whole genome of viruses from samples like bat feces (Li et al 2019). In this case, without the use of any other methods, and after using so old fecal swab or fecal swab RNA with no bacteria in it, how did they recover such good quality viral reads?

6. The assembly method and the actual assembly accession for RaTG13 is not described or linked to MN669532 and also no assembly method is specified in the raw data SRX7724752 and the Illumina run. Therefore, no assembly data is available for RaTG13 genome.

7. After blasting the RaTG13 genome against the SRA, ~1700 reads can be retrieved which covers only 252 Kb of the total 3.3 Gb. The genome size of RaTG13 is known to be ~30 kb. Therefore this is ~8x coverage, which is quite less and insufficient to arrive to a definitive assembly. Then how was the sequence MN669532 used so confidently by various researchers without any doubt?

8. We also compared the fecal/anal swab from the same species, i.e. *Rhinolophus affinis* (Fig.2) and fecal swab from another bat (Fig. 3) and it clearly shows that the other two swabs showed normal findings, with 70-90% bacterial reads and very few reads associated with the host. Also these swabs do not show sequences coming from other animals.

9. Similar findings have been documented in a latest preprint by Zhang, D. (Zhang, 2020) https://zenodo.org/record/3969272#.Xypwfn5S-Un.

doi:10.20944/preprints202008.0205.v1

Problems in the Amplicon sequencing data:

We found that some amplicon sequencing data for RaTG13 (SRX8357956) was submitted in May 2020.

1. No indications of amplicon sequencing given by Zhou et al 2020 about the amplicon sequencing of RaTG13. There are in total 33 spots with forward and reverse sequences.

2. This sequencing shows that the dates are 2017 and 2018. However, the submission has been done in 2020. This sequencing has never been mentioned in any publications. Also, it does not cover the entire genome and major gaps are seen in various regions.

3. There are two contrasting sequences for a single patch (spots 23 and spot 24), e.g. shows 94-96% similarity to that of MN669532.1. However, another spot the same sequence showed 99% similarity to the described RaTG13 consensus MN669532.1.

4. In general, the amplicons show 97-99% similarity with the MN669532.1. However, it does not cover the entire genome and major gaps are seen in various regions.

5. Also the RdRp derived from the amplicon sequencing is incomplete and does not match with RdRp of 4991 KP876546.1. Around 170 bases from 370 base sequences are missing and it shows 2 base mismatches.

Conclusions:

a. Our main objection is that the fecal swab from which RaTG13 sequence is derived does not appear like a normal fecal sample due to the above listed things.

b. RaTG13 sequence has been used extensively for all genomic comparisons as it is believed to be the next relative of SARS-CoV-2.

c. However, the nature of the fecal swab appears very suspicious, with 70% of eukaryotic sequences also from sources which should not have been detected in bat feces like mexican bats, squirrels, flying foxes, red foxes, etc.).

d. And most importantly, there is negligible abundance of bacteria. Bacteria constitute a major part of any feces, irrespective if it is an animal or bird or any eukaryote.

e. The reads from which the viral sequence of RaTG13 was derived appears not to be affected. An almost complete assembly is assumed to be had been built from this raw data (Illumina reads). How did so good data come from an otherwise abnormal looking, old and degraded fecal swab sample preserved for 7-8 years?

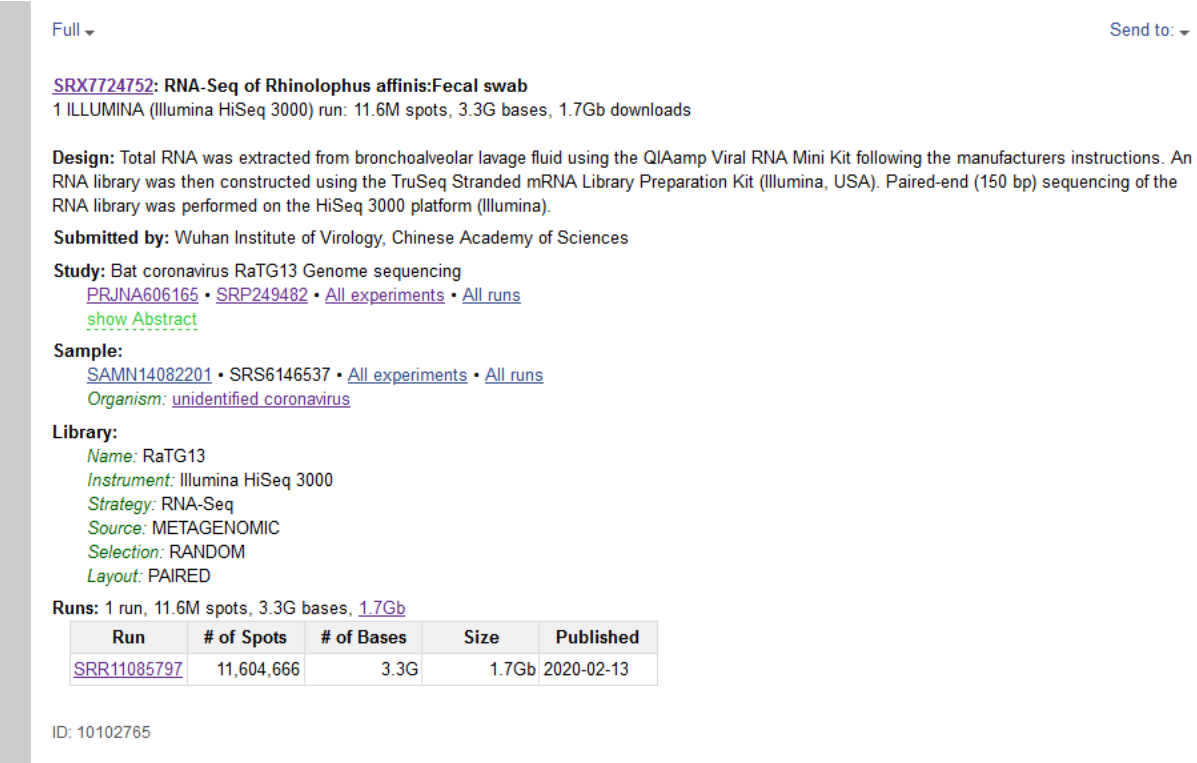
f. The amplicon data is incomplete and submitted much later and undescribed anywhere.

g. The question is why are these anomalies? And if these are there, should the scientific community really rely on the RaTG13 genome sequence MN996532.1? Should this data be used for further important experiments?

doi:10.20944/preprints202008.0205.v1

Figures:

Fig.1 RNA-Seq of Rhinolophus affinis:Fecal swabTaxonomy Analysis (RaTG13)



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

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

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

Study: Bat coronavirus RaTG13 Genome sequencing
PRJNA606165 • SRP249482 • All experiments • All runs
show Abstract

Sample:
SAMN14082201 • SRS6146537 • All experiments • All runs
Organism: unidentified coronavirus

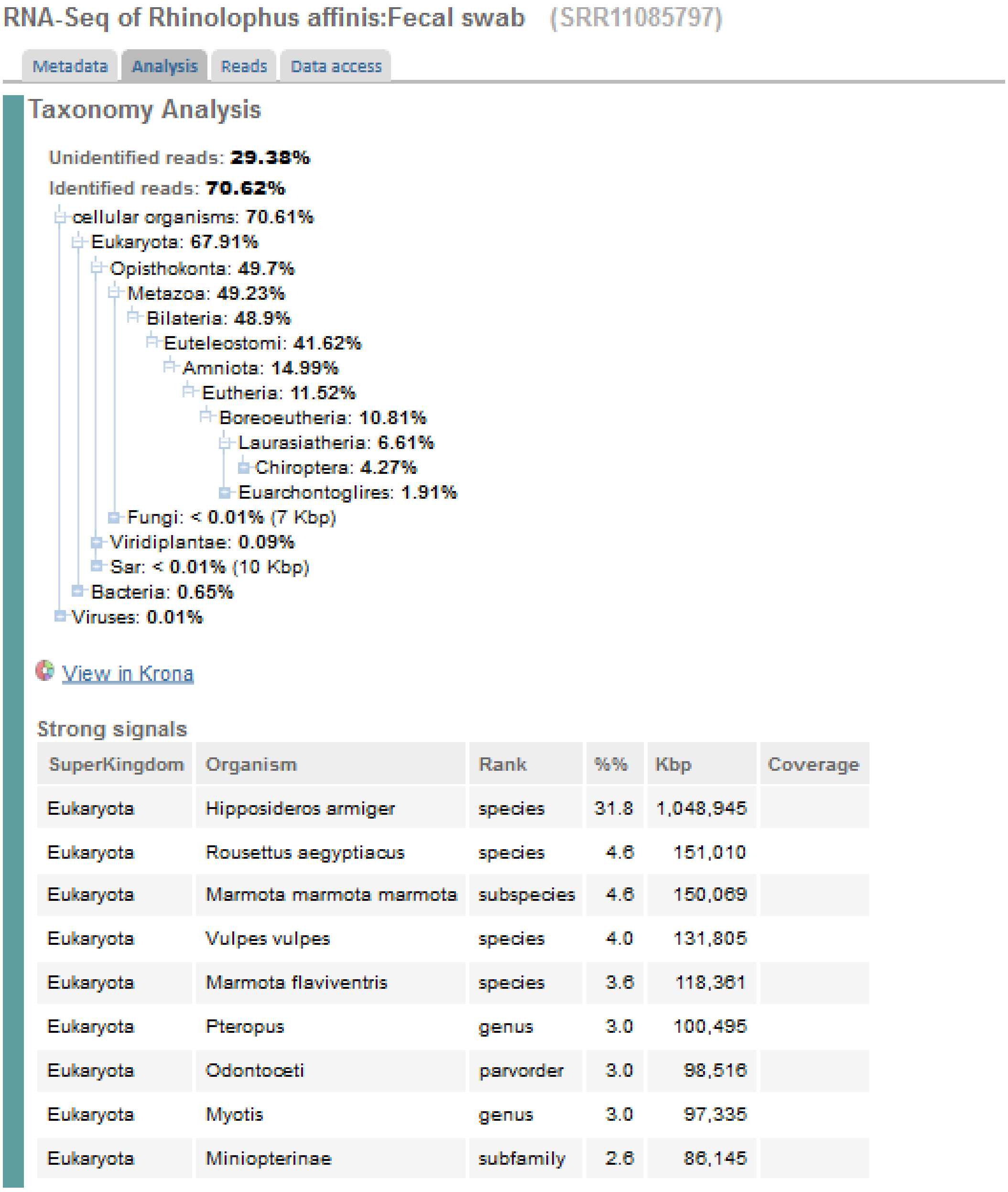
Library:
Name: RaTG13
Instrument: Illumina HiSeq 3000
Strategy: RNA-Seq
Source: METAGENOMIC
Selection: RANDOM
Layout: PAIRED

Runs: 1 run, 11.6M spots, 3.3G bases, 1.7Gb

Run	# of Spots	# of Bases	Size	Published
SRR11085797	11,604,666	3.3G	1.7Gb	2020-02-13

ID: 10102765

Fig1a. RNA-Seq of *Rhinolophus affinis*:Fecal swab **(RaTG13)**



RNA-Seq of Rhinolophus affinis:Fecal swab (SRR11085797)

Taxonomy Analysis

Unidentified reads: **29.38%**

Identified reads: **70.62%**

- cellular organisms: **70.61%**
  - Eukaryota: **67.91%**
    - Opisthokonta: **49.7%**
      - Metazoa: **49.23%**
        - Bilateria: **48.9%**
          - Euteleostomi: **41.62%**
            - Amniota: **14.99%**
              - Eutheria: **11.52%**
                - Boreoeutheria: **10.81%**
                  - Laurasiatheria: **6.61%**
                    - Chiroptera: **4.27%**
                  - Euarchontoglires: **1.91%**
      - Fungi: **< 0.01%** (7 Kbp)
    - Viridiplantae: **0.09%**
    - Sar: **< 0.01%** (10 Kbp)
  - Bacteria: **0.65%**
- Viruses: **0.01%**

View in Krona

Strong signals

SuperKingdom	Organism	Rank	%%	Kbp	Coverage
Eukaryota	Hipposideros armiger	species	31.8	1,048,945	
Eukaryota	Rousettus aegyptiacus	species	4.6	151,010	
Eukaryota	Marmota marmota marmota	subspecies	4.6	150,069	
Eukaryota	Vulpes vulpes	species	4.0	131,805	
Eukaryota	Marmota flaviventris	species	3.6	118,361	
Eukaryota	Pteropus	genus	3.0	100,495	
Eukaryota	Odontoceti	parvorder	3.0	98,516	
Eukaryota	Myotis	genus	3.0	97,335	
Eukaryota	Miniopterinae	subfamily	2.6	86,145	

Fig. 1b. Distribution of the reads in the raw data. The individual distribution is given and in the second part, the reads which contribute to a higher extent are given.

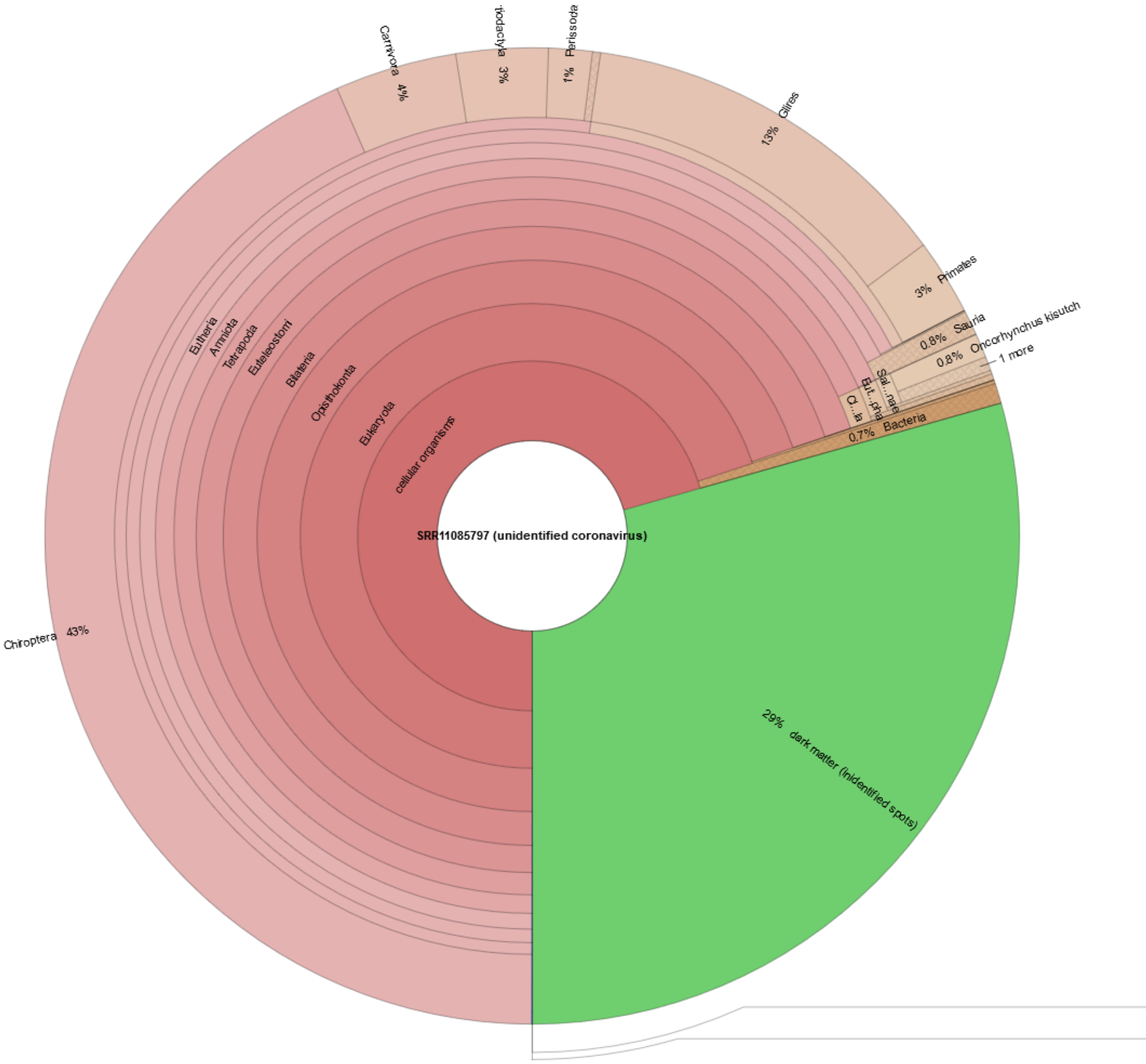
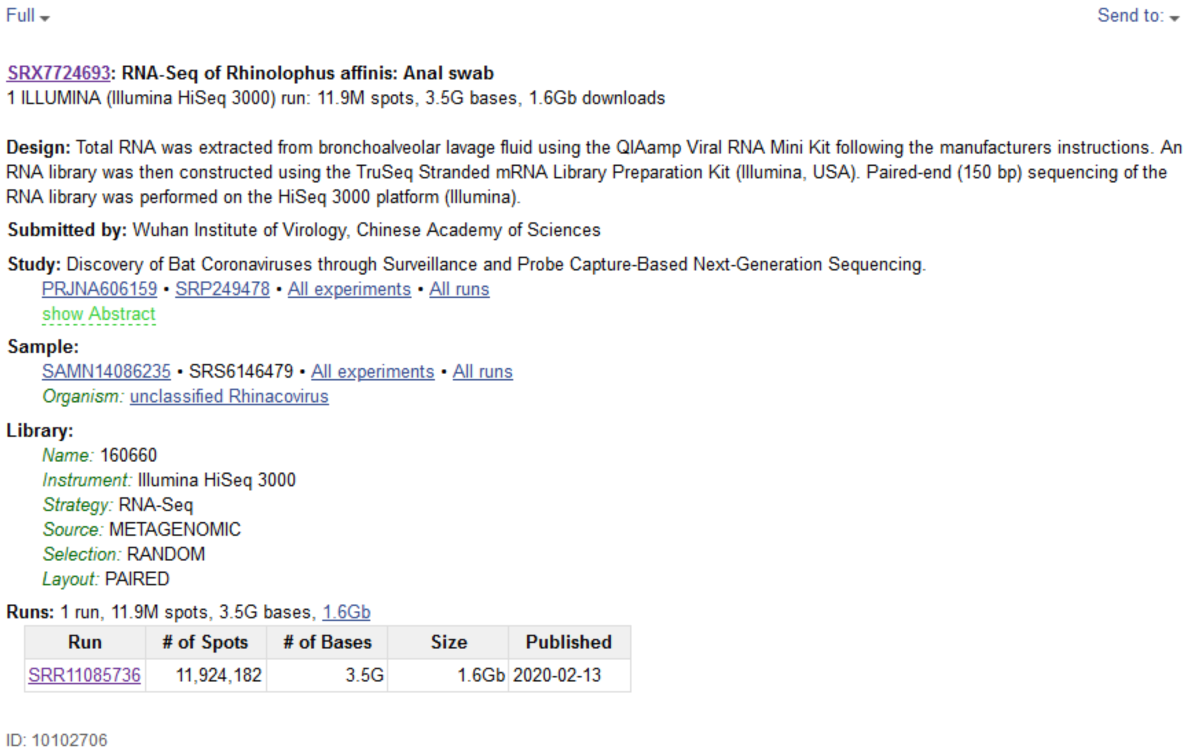


Fig.1 c. Krona chart of RaTG13 raw data, 29% unidentified reads, 43% Chiroptera, 13% Gileres, 3% Primates, 0.7% bacteria and 0.024% RaTG13 reads

Fig 2. RNA-Seq of Rhinolophus affinis: Fecal swab Taxonomy Analysis

https://www.ncbi.nlm.nih.gov/sra/SRX7724693[accn]



SRX7724693: RNA-Seq of Rhinolophus affinis: Anal swab
1 ILLUMINA (Illumina HiSeq 3000) run: 11.9M spots, 3.5G bases, 1.6Gb downloads

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

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

Study: Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing.
PRJNA606159 • SRP249478 • All experiments • All runs
show Abstract

Sample:
SAMN14086235 • SRS6146479 • All experiments • All runs
Organism: unclassified Rhinacovirus

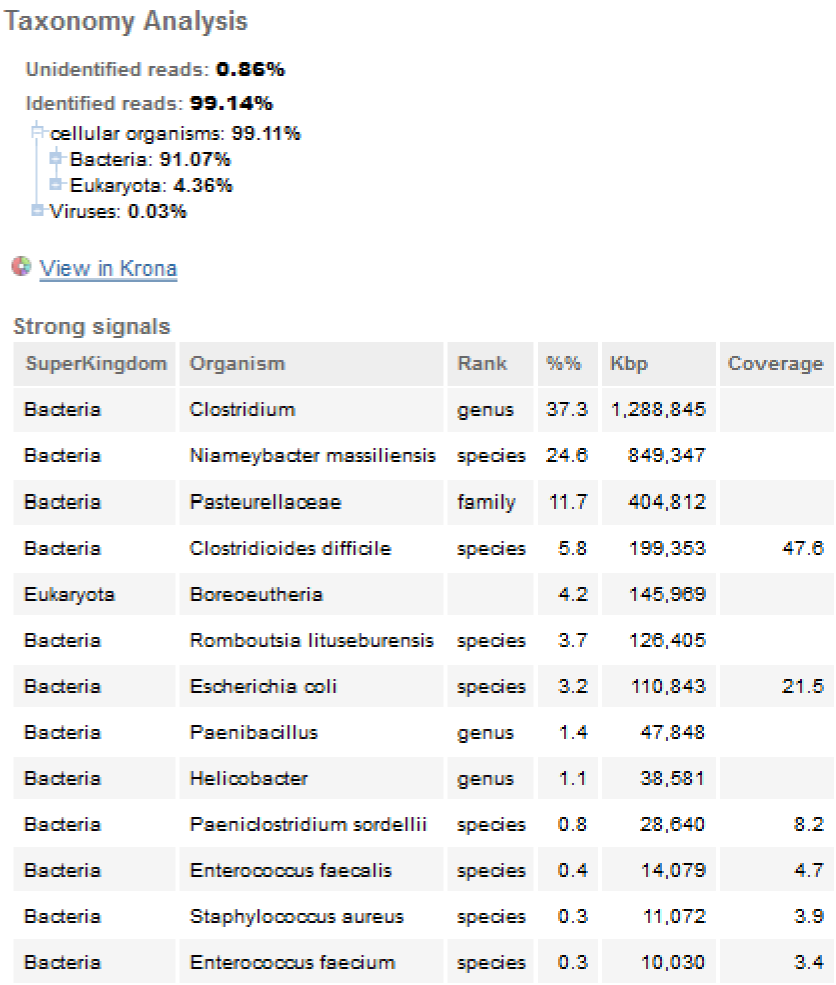
Library:
Name: 160660
Instrument: Illumina HiSeq 3000
Strategy: RNA-Seq
Source: METAGENOMIC
Selection: RANDOM
Layout: PAIRED

Runs: 1 run, 11.9M spots, 3.5G bases, 1.6Gb

Run	# of Spots	# of Bases	Size	Published
SRR11085736	11,924,182	3.5G	1.6Gb	2020-02-13

ID: 10102706

Fig. 2a. RNA-Seq of Rhinolophus affinis: Anal swab (SRR11085736)



Taxonomy Analysis

Unidentified reads: **0.86%**

Identified reads: **99.14%**

- cellular organisms: 99.11%
  - Bacteria: 91.07%
  - Eukaryota: 4.36%
- Viruses: 0.03%

View in Krona

Strong signals

SuperKingdom	Organism	Rank	%%	Kbp	Coverage
Bacteria	Clostridium	genus	37.3	1,288,845	
Bacteria	Niameybacter massiliensis	species	24.6	849,347	
Bacteria	Pasteurellaceae	family	11.7	404,812	
Bacteria	Clostridioides difficile	species	5.8	199,353	47.6
Eukaryota	Boreoeutheria		4.2	145,969	
Bacteria	Romboutsia lituseburensis	species	3.7	126,405	
Bacteria	Escherichia coli	species	3.2	110,843	21.5
Bacteria	Paenibacillus	genus	1.4	47,848	
Bacteria	Helicobacter	genus	1.1	38,581	
Bacteria	Paeniclostridium sordellii	species	0.8	28,640	8.2
Bacteria	Enterococcus faecalis	species	0.4	14,079	4.7
Bacteria	Staphylococcus aureus	species	0.3	11,072	3.9
Bacteria	Enterococcus faecium	species	0.3	10,030	3.4

Fig. 2b. Distribution of the reads in the raw data. The individual distribution is given and in the second part, the reads which contribute to a higher extent are given.

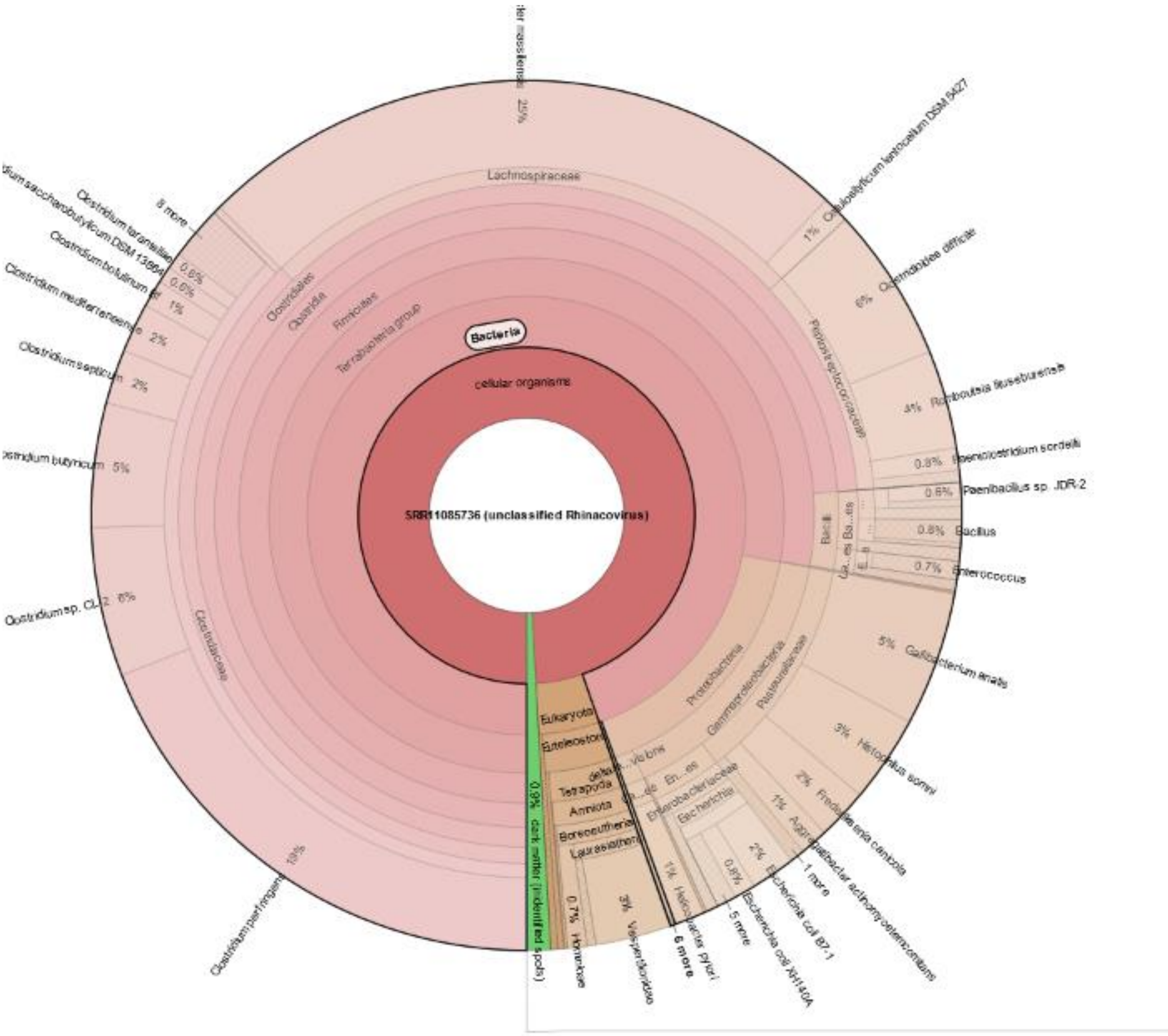
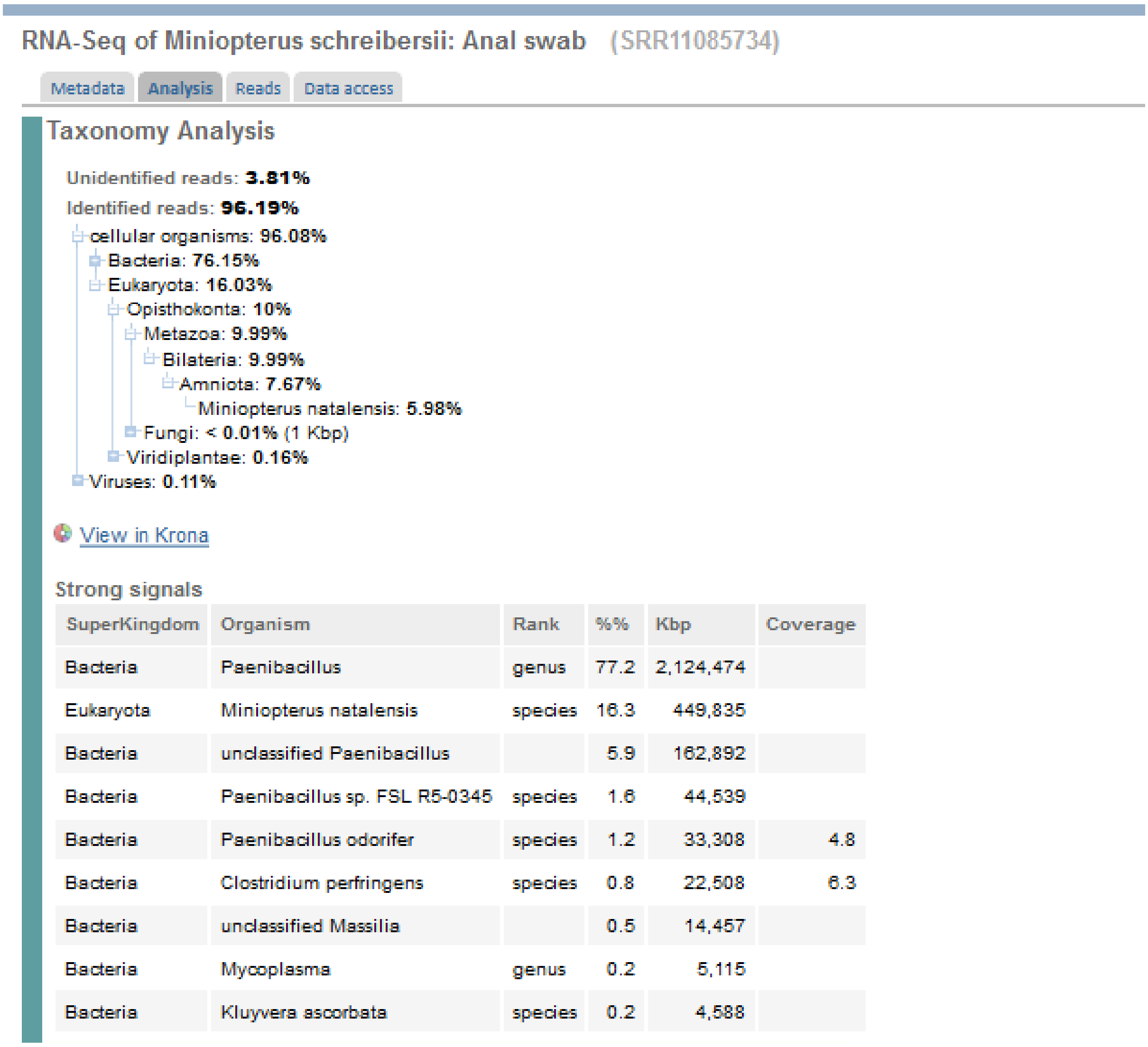


Fig. 2c. Krona chart of the anal swab of Rhinolophus affinis: Fecal swab Taxonomy

doi:10.20944/preprints202008.0205.v1

Fig 3 RNA-Seq of *Miniopterus schreibersii*: Fecal swab Taxonomy Analysis



RNA-Seq of Miniopterus schreibersii: Anal swab (SRR11085734)

Taxonomy Analysis

Unidentified reads: **3.81%**

Identified reads: **96.19%**

- cellular organisms: 96.08%
  - Bacteria: 76.15%
  - Eukaryota: 16.03%
    - Opisthokonta: 10%
      - Metazoa: 9.99%
        - Bilateria: 9.99%
          - Amniota: 7.67%
            - Miniopterus natalensis: 5.98%
      - Fungi: < 0.01% (1 Kbp)
    - Viridiplantae: 0.16%
- Viruses: 0.11%

View in Krona

Strong signals

SuperKingdom	Organism	Rank	%%	Kbp	Coverage
Bacteria	Paenibacillus	genus	77.2	2,124,474	
Eukaryota	Miniopterus natalensis	species	16.3	449,835	
Bacteria	unclassified Paenibacillus		5.9	162,892	
Bacteria	Paenibacillus sp. FSL R5-0345	species	1.6	44,539	
Bacteria	Paenibacillus odorifer	species	1.2	33,308	4.8
Bacteria	Clostridium perfringens	species	0.8	22,508	6.3
Bacteria	unclassified Massilia		0.5	14,457	
Bacteria	Mycoplasma	genus	0.2	5,115	
Bacteria	Kluyvera ascorbata	species	0.2	4,588	

Fig. 3a. RNA-Seq of fecal swab *Miniopterus schreibersii*

doi:10.20944/preprints202008.0205.v1

Fig. 3c. Krona chart of *Miniopterus schreibersii*: Fecal swab Taxonomy

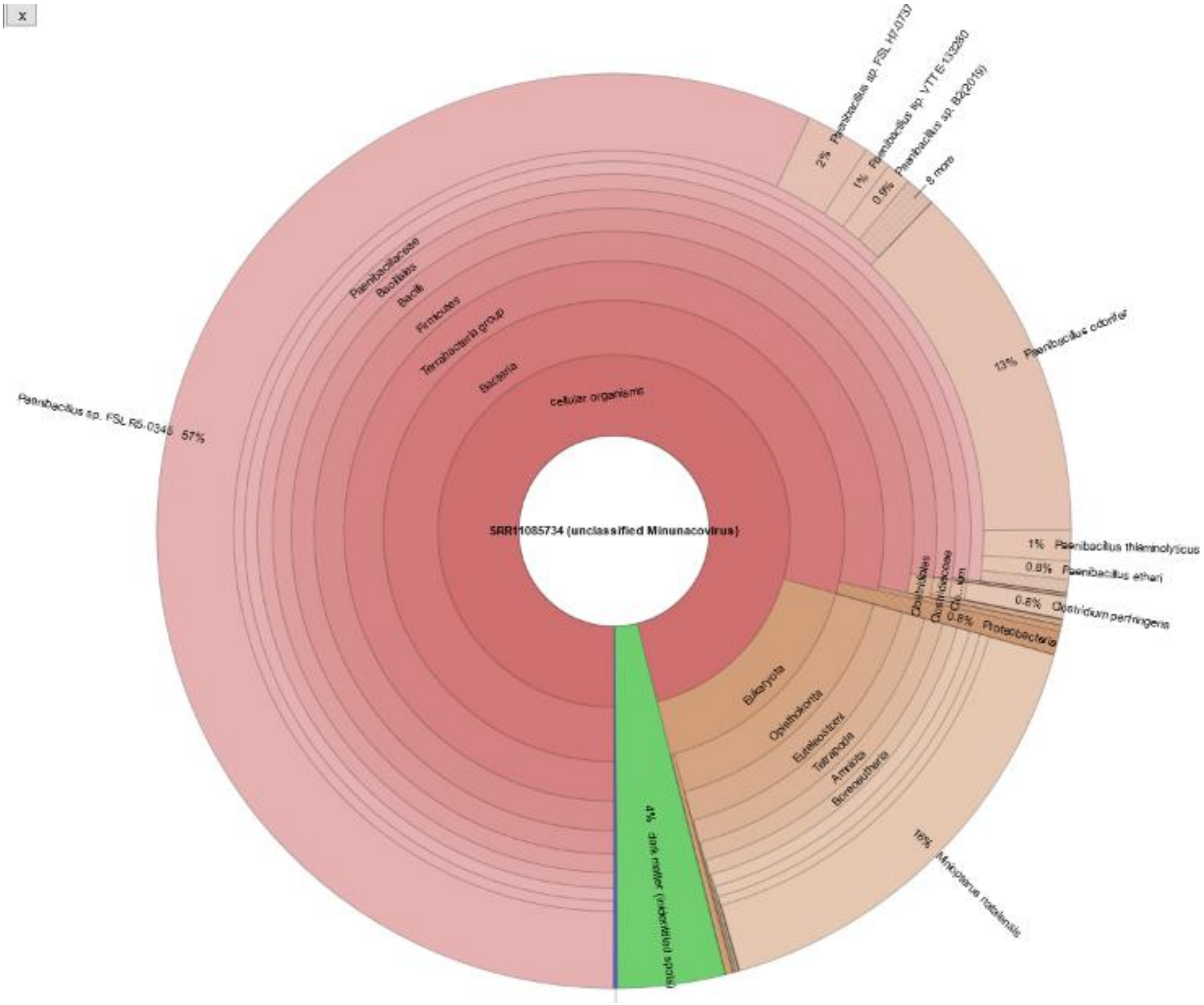
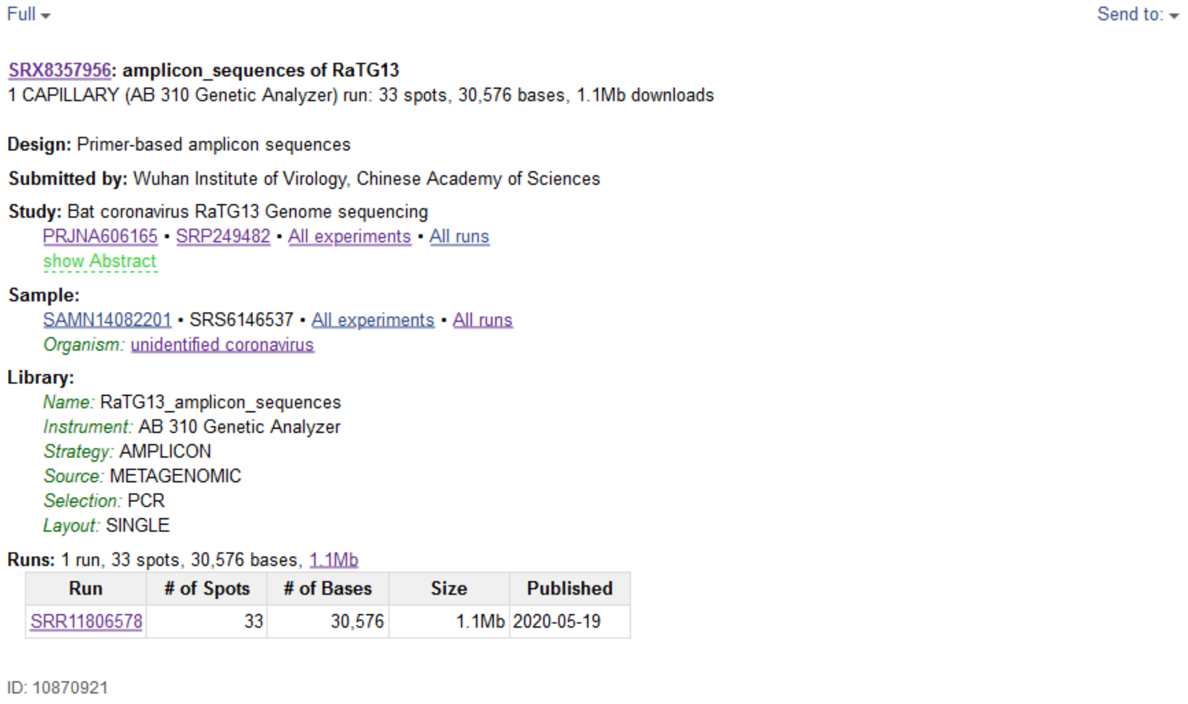


Fig. 4



SRX8357956: amplicon_sequences of RaTG13
1 CAPILLARY (AB 310 Genetic Analyzer) run: 33 spots, 30,576 bases, 1.1Mb downloads

Design: Primer-based amplicon sequences

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

Study: Bat coronavirus RaTG13 Genome sequencing
PRJNA606165 • SRP249482 • All experiments • All runs
show Abstract

Sample:
SAMN14082201 • SRS6146537 • All experiments • All runs
Organism: unidentified coronavirus

Library:
Name: RaTG13_amplicon_sequences
Instrument: AB 310 Genetic Analyzer
Strategy: AMPLICON
Source: METAGENOMIC
Selection: PCR
Layout: SINGLE

Runs: 1 run, 33 spots, 30,576 bases, 1.1Mb

Run	# of Spots	# of Bases	Size	Published
SRR11806578	33	30,576	1.1Mb	2020-05-19

ID: 10870921

References:

2020. https://www.thetimes.co.uk/article/seven year covid trail revealed l5vxt7jqp. *The Sunday Times.*

Ge, X. Y., Wang, N., Zhang, W., Hu, B., Li, B., Zhang, Y. Z., Zhou, J. H., Luo, C. M., Yang, X. L., Wu, L. J., Wang, B., Zhang, Y., Li, Z. X. & Shi, Z. L. 2016. Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virol. Sin.,* 31**,** 31-40.

Rahalkar, Monali C. & Bahulikar, Rahul A. 2020. Understanding the origin of 'BatCoVRaTG13', a virus closest to SARS-CoV-2.

Zhang, Daoyu 2020. Anomalies in BatCoV/RaTG13 sequencing and provenance. *https://zenodo.org/record/3969272#.Xy0m5jVS_IX*.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. D., Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., Zheng, X. S., Zhao, K., Chen, Q. J., Deng, F., Liu, L. L., Yan, B., Zhan, F. X., Wang, Y. Y., Xiao, G. F. & Shi, Z. L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* 579**,** 270-273.