Short Note

The anomalous nature of the fecal swab sample used for RaTG13 genome assembly as revealed by NGS data analysis

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, a SARS-like beta coronavirus, which exists in the form of a genome sequence, is the closest relative of SARS-CoV-2 reported till date. The sample from which RaTG13 virus was sequenced was a bat fecal swab collected in 2013 from Tongguan, Mojiang, Yunnan province, China. The genome data for RaTG13, MN996532.1, was deposited on 27th Jan 2020 and the raw data (Illumina reads) was deposited a fortnight later on 13th Feb 2020 https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn]. Comparison of the RNA Seq data of RaTG13 fecal swab sample to the corresponding data from the bat fecal swabs deposited by the same working group indicated that the raw data seemed to be anomalous in several aspects. Thirty percent of the reads did not match with anything. From the rest of the 70%, an abnormal high proportion was contributed by reads derived from eukaryotes (~68%). These matched with the sequences of not one but various bat species (round leaf bats, fruit bats and other bats) and animal species (squirrels, foxes, etc.) as per Krona analysis included with the SRA data. The proportion of the bacterial reads in the swab was exceptionally low, i.e. 0.7%, which is abnormal, compared to the 70-90% bacterial abundance in other bat fecal swabs. Furthermore, we also found another set of raw data associated with RaTG13, amplicon sequencing of the genome (SRX8357956), which was submitted in May 2020. Analysis of the amplicons by BLAST showed that these collectively do not cover the whole genome (MN996532.1). On closer inspection, the dates mentioned in the files of the sequenced amplicons were also found to be older (2017, 2018). Collectively, the anomalies in the raw data of RaTG13 pose an important question about the overall authenticity of the RaTG13 genome sequence.

Key words: RaTG13; SARS-COV-2; Illumina sequencing, amplicon sequencing, NGS; fecal swab

Covid-19 has been a devastating pandemic affecting more than nineteen million people and killing about three quarter million people till date (8th August 2020). SARS-CoV2, the virus responsible for the pandemic is most similar to RaTG13 (a bat derived coronavirus) on the genomic level. RaTG13 has been known as the sister virus of SARS-CoV-2 as its shows the closest overall genomic similarity (96.2%) to SARS-CoV-2 genome (Zhou et al., 2020). RaTG13 has been used for various comparative experiments with SARS-CoV-2. These include: the capacity of its spike to bind to human ACE-2 and the capacity to cause human infections (Wrobel et al., 2020), the evolutionary analysis and prediction of a common ancestor of SARS-CoV-2 and RaTg13 (Boni et al., 2020), and many more upcoming papers.

RaTG13 is a beta SARS-like corona virus and the name was introduced to us in 2020 (Zhou et al., 2020). The sequence of RaTG13 was retrieved by RNA sequencing using a next generation sequencing approach, though the sample was collected in 2013 (Zhou et al., 2020). The RNA sample was that of a bat fecal swab collected in July 2013, from Tongguan mineshaft in Yunnan. The details of the location were predicted earlier (Arbuthnott et al., 2020, Rahalkar and Bahulikar, 2020). However, in a recent reply to the Science magazine it has been clarified by Dr. Zheng-Li Shi that the TG in RaTG13 is for Tongguan, Mojiang, Yunna, China (Cohen, 2020). She has also confirmed that the old name of RaTG13 virus was BtCoV/CoV4991, described earlier (Ge et al., 2016). However, the sample appears to be finished or disintegrated or not available to the scientific community as per a recent media investigation (Arbuthnott et al., 2020). Also, Zheng-Li Shi has confirmed that her lab has never cultured this virus and it is not in live condition in her lab (Cohen, 2020). Therefore, the entire scientific community has to rely on the RaTG13 genome available. And if the RaTG13 genome sequence is to be used in all the bioinformatics and model experiments, the pre-supposition is that the sequence of this virus should be accurate and based on a reliable raw data. Therefore, we have looked at the RaTG13 raw data in a closer manner in this paper.

The name RaTG13 was first mentioned in 2020 (Zhou et al., 2020) and the full genome sequence was not available before January 2020 on any of the databases, to the best of our knowledge. The Illumina based NGS sequence of RaTG13 **MN996532.1** was deposited on 27th Jan 2020 and the raw data was available a little later on 13th Feb 2020 https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn].

The older name of RaTG13 was BtCoV/4991, where the names were based on RdRP sequences and sample numbers (Cohen, 2020). A 370 base RdRp fragment of BtCoV/4991 (KP378696.1) or RaTG13 shows the highest similarity to SARS-CoV-2 with only 3-5 bases difference (blast comparison with the SARS-CoV-2 sequences deposited till date). Also, BtCoV/4991 or RaTG13 has a great significance as it had been recovered from the same location, a mineshaft in Tongguan, where six miners were afflicted with a suspiciously COVID-19 like pneumonia in 2012, and three succumbed to the infection and died (Arbuthnott et al., 2020, Rahalkar and Bahulikar, 2020). Thus, BtCoV/4991 or RaTG13 is also the first and the only beta SARS-like CoV known so far associated with Tongguan mineshaft where lethal human pneumonia cases were reported in 2012 (Arbuthnott et al., 2020, 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 RaTG13 fecal swab Illumina data https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn]:

1. The genome of RaTG13 (MN996532.1) is derived from a fecal swab sample collected in 2013 as per the description. However, the Illumina sequencing entry of SRX7724752, the sample is recorded as being extracted from a BAL fluid (broncho alveolar lavage) (Fig. 1a).

2. The total raw data is 3.3 gigabases (Fig. 1b). 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% is seen to be contributed by eukaryotic sequences (Fig. 1b). This is highly unusual as it is a fecal swab and the analysis of other bat fecal or anal swabs usually do not show such high proportion of eukaryotic RNA.

3. Within the 68% of the eukaryotic sequences, the bat sequences are about 36-40% (Fig 1a.). The rest of the 30% sequences are contributed by squirrels (Marmota), flying foxes, foxes, and other types of animals (Fig.1 b). In a bat fecal swab, the majority of the eukaryotic RNA should be arising from the same species, e.g. in this case it should be of *Rhinoplohus affinis*. Given the fact that *Rhinolophus affinis* bat has not been sequenced completely, it could show similarities with other bat species, such as *Hipposideros* (round leaf bats) or *Rousettus* (egyptian fruit bats). However, the rest of the hits are from totally unrelated taxa, such as *Marmota* (squirrels), *Vulpus* (foxes), *Pteropus* (mega-bats), etc. Similar discrepancies in the raw data have been pointed out recently (Zhang, 2020).

4. Another major discrepancy is that the RNA sequencing data shows extremely less abundance of bacteria, only 0.65%. This is far too less in comparison with other fecal or anal swab of bats, which show a very high proportion of bacterial sequences ~76-90% (Fig.2 and Fig.3). SRA data of six other bat fecal swabs submitted by the same group also showed a high abundance of bacterial reads (details not shown). Bacteria are usually the highest constituents of gut flora and hence contribute to a high extent to a fecal sample.

5. The coronavirus sequence (RaTG13) contributes to ~0.003% of the total sequence reads. These raw reads were used to build an almost complete assembly, though the overall coverage is less ~7-8X. Though there were less overlaps in a few regions, there are only 3 gaps as per our analysis using RaTG13 as the reference genome. 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 a seven year old fecal swab or fecal swab RNA it is surprising that how the viral reads were of a better quality.

6. The assembly method and the actual assembly accession for RaTG13 is not described or linked to the whole genome of RaTG13, i.e. MN669532.1 and also no assembly method is specified in the raw data SRX7724752. Also, no assembly data accession number is available for RaTG13 genome as per our information and searches.

7. After blasting the RaTG13 genome against the SRA, ~1700 reads can be retrieved which covers only a small portion, i.e. 252 kb of the total 3.3 Gbases. The genome size of RaTG13 is known to be ~30 kb. According to our knowledge, this is ~8x coverage is low and may be insufficient to arrive at a definitive genome assembly.

8. We also compared the fecal/anal swab RNA Seq data deposited for the same bat species, i.e. *Rhinolophus affinis* (Fig.2) and fecal swab from another bat (Fig. 3). It is clearly seen 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 affiliated with other animals.

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

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 has been given by Zhou et al 2020, or in any of the recent publications by the WIV workgroup.

2. A total of 33 spots are present in the raw data (Fig.4). The sequencing file names indicate that the dates are from 2017 and 2018. However, the submission has been done in May 2020.

3. There are two contrasting sequences for a single patch (spot 23 and spot 24), e.g. shows 94-96% similarity to that of MN669532.1. However, two spots (22 and 25) covering the same area showed 99% similarity to the described RaTG13 consensus MN669532.1.

4. In general, most of the amplicons showed 97-99% similarity with that of MN669532.1. However, collectively, the spots do not cover the entire genome and major gaps are seen in various regions.

5. RdRp derived from the amplicon sequencing is incomplete (spots 31 and 32) and does not match with RdRp of BtCoV/4991 KP876546.1. Around 170 bases from 370 base sequences are missing and it shows 2 base mismatches.

Conclusions:

a. Our main grievance is that the fecal swab from which RaTG13 sequence has been derived appears as an anomalous fecal swab as pointed above with respect to its community composition. The swab shows 70% of eukaryotic sequences from sources which should not have been detected in *Rhinolophus* bat feces such as mexican bats, squirrels, flying foxes, red foxes, etc. And most importantly, there is extreme low representation of bacteria. Bacteria constitute a major part of feces from any eukaryotic organism.

b. RaTG13 genome sequence has been used extensively for in various evolutionary calculations, simulation experiments and would be used in future for bioinformatics and experimental comparisons. And therefore, all the data associated with RaTG13 should be inspected properly.

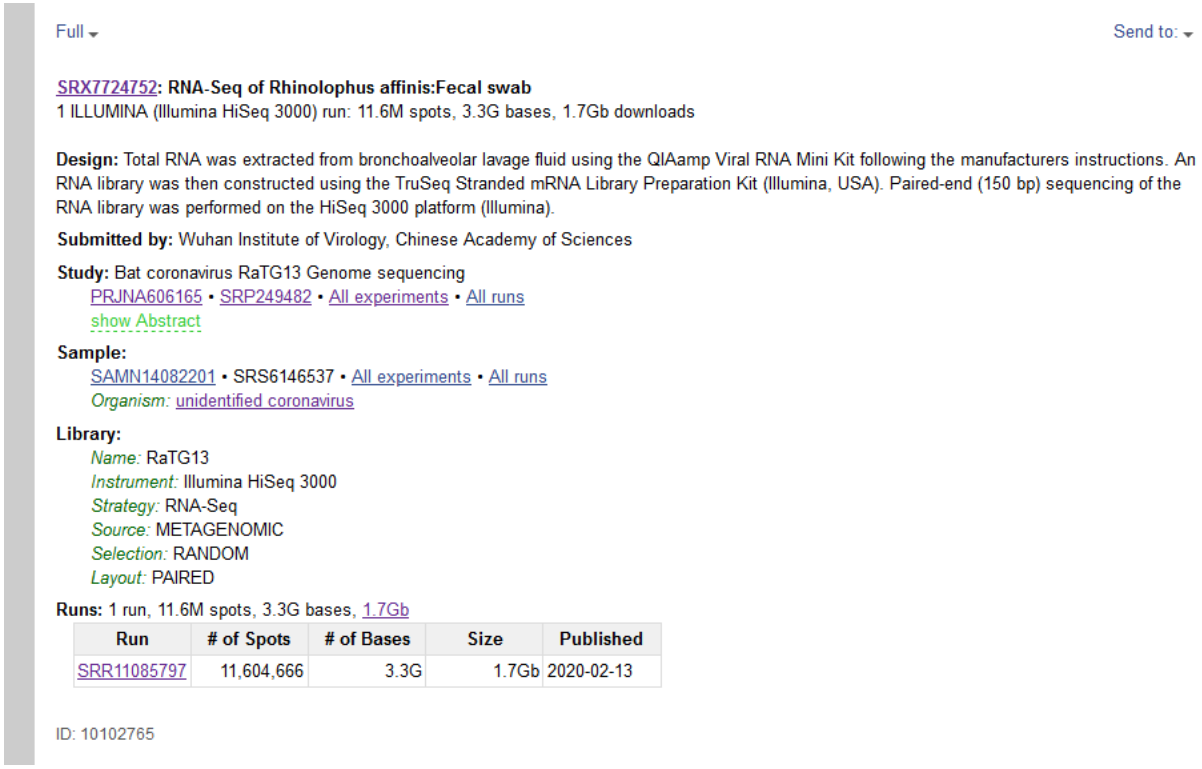
c. The reads from which the viral sequence of RaTG13 was assembled appears not to be affected by the anomalous nature of the RNA Seq data of the fecal swab. An almost complete assembly is assumed to have been built from this raw data (Illumina reads). An important question is how did this considerably good data related to the virus come from this swab? And if it was not degraded, what are the reasons of its anomalous composition?

d. The amplicon data is incomplete and does not help us in further confirming the RaTG13 whole genome sequence.

e. Considering the anomalous nature of the raw data presented for RaTG13 (both Illumina and amplicon sequence) it would be a real question can the scientific community rely on the integrity of the RaTG13 genome sequence MN996532.1? Moreover, with the discrepancies pointed out in the raw data of RaTG13, we suggest that RaTG13 genome sequence should be interpreted with caution.

Figures:

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



Full ▾ Send to: ▾

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

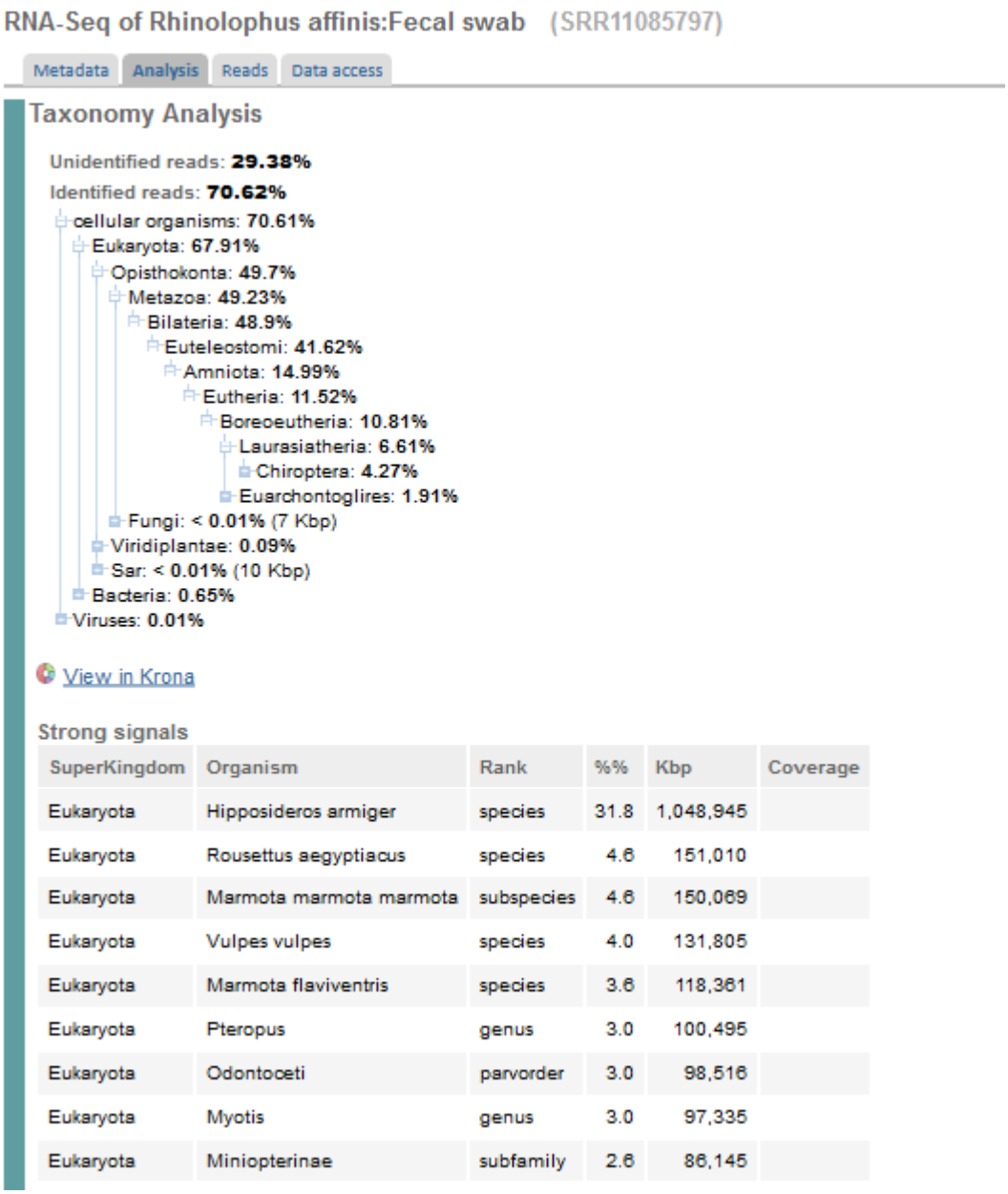
Library:
Name: RaTG13
Instrument: Illumina HiSeq 3000
Strategy: RNA-Seq
Source: METAGENOMIC
Selection: RANDOM
Layout: PAIRED

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)

Metadata | Analysis | Reads | Data access

Taxonomy Analysis

Unidentified reads: **29.38%**

Identified reads: **70.62%**

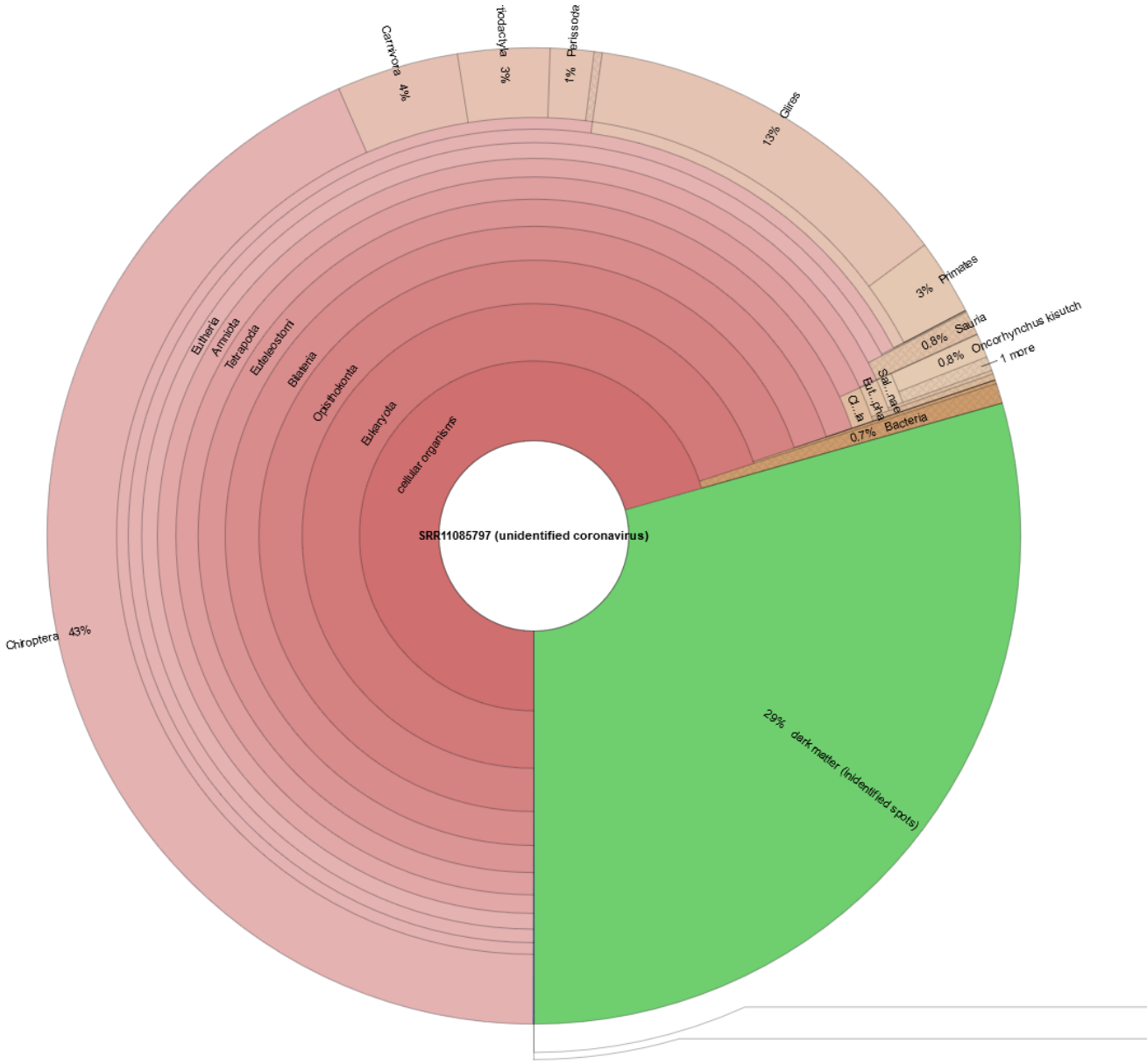
- cellular organisms: **70.61%**
  - Eukaryota: **67.91%**
    - Opisthokonta: **49.7%**
      - Metazoa: **49.23%**
        - Bilateria: **48.9%**
          - Euteleostomi: **41.62%**
            - Amniota: **14.99%**
              - Eutheria: **11.52%**
                - Boreoeutheria: **10.81%**
                  - 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 with strong signals, i.e. which contribute to a higher extent are given in descreasing order.



186

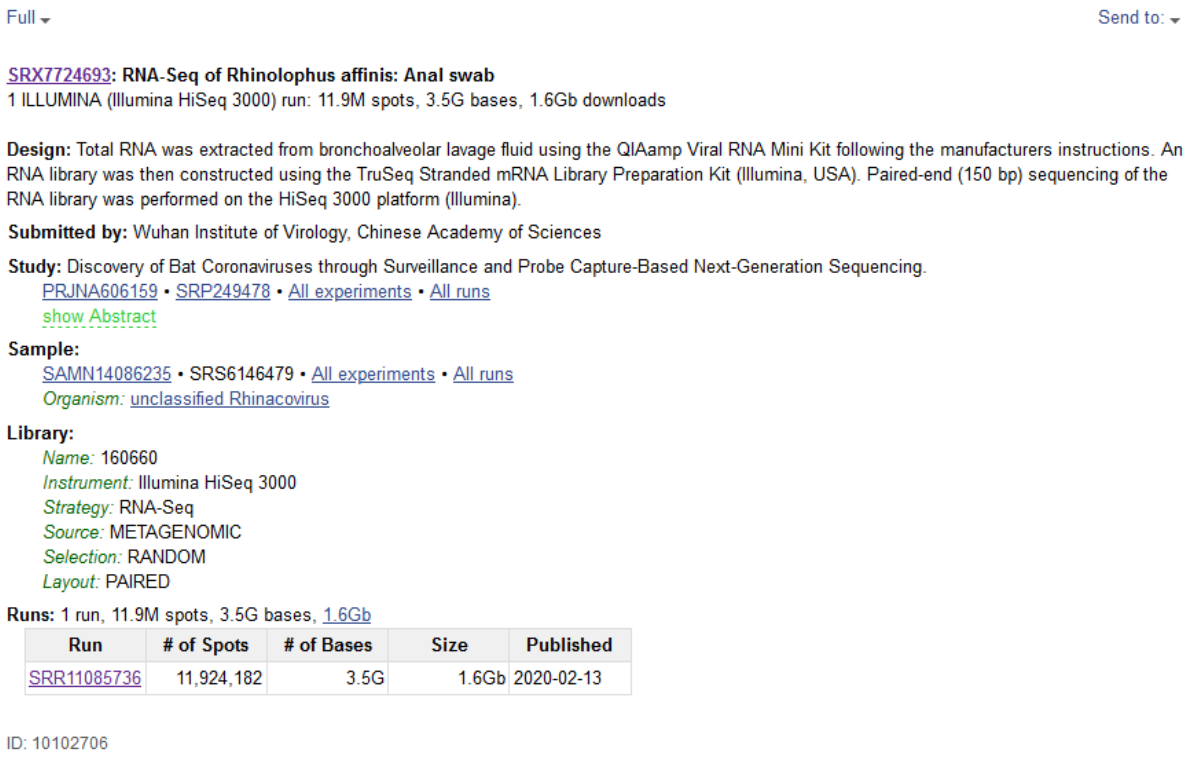
187

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

190

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

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



Full ▾ Send to: ▾

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

193

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

195

Taxonomy Analysis

Unidentified reads: **0.86%**

Identified reads: **99.14%**

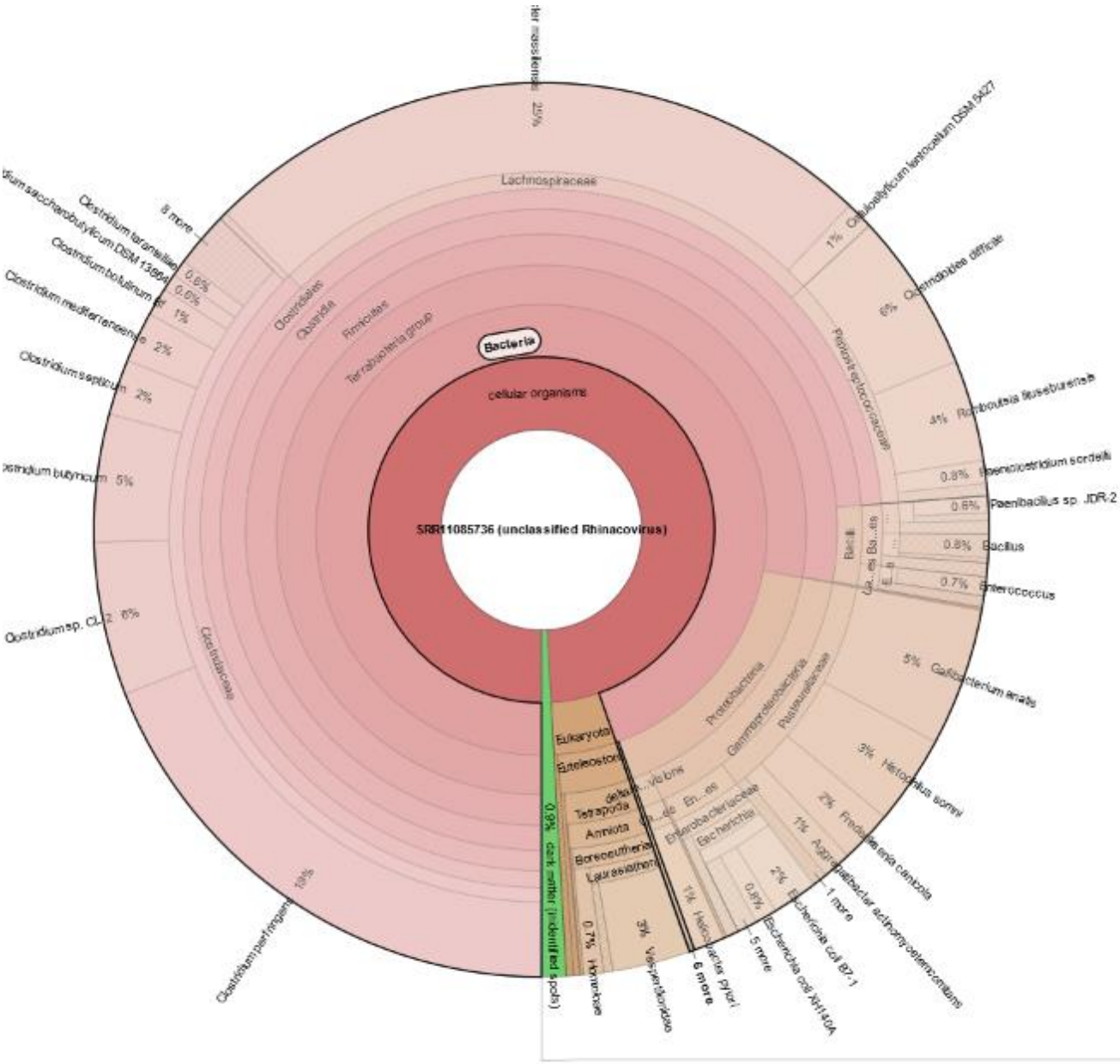
- cellular organisms: 99.11%
  - Bacteria: 91.07%
  - Eukaryota: 4.36%
- Viruses: 0.03%

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.



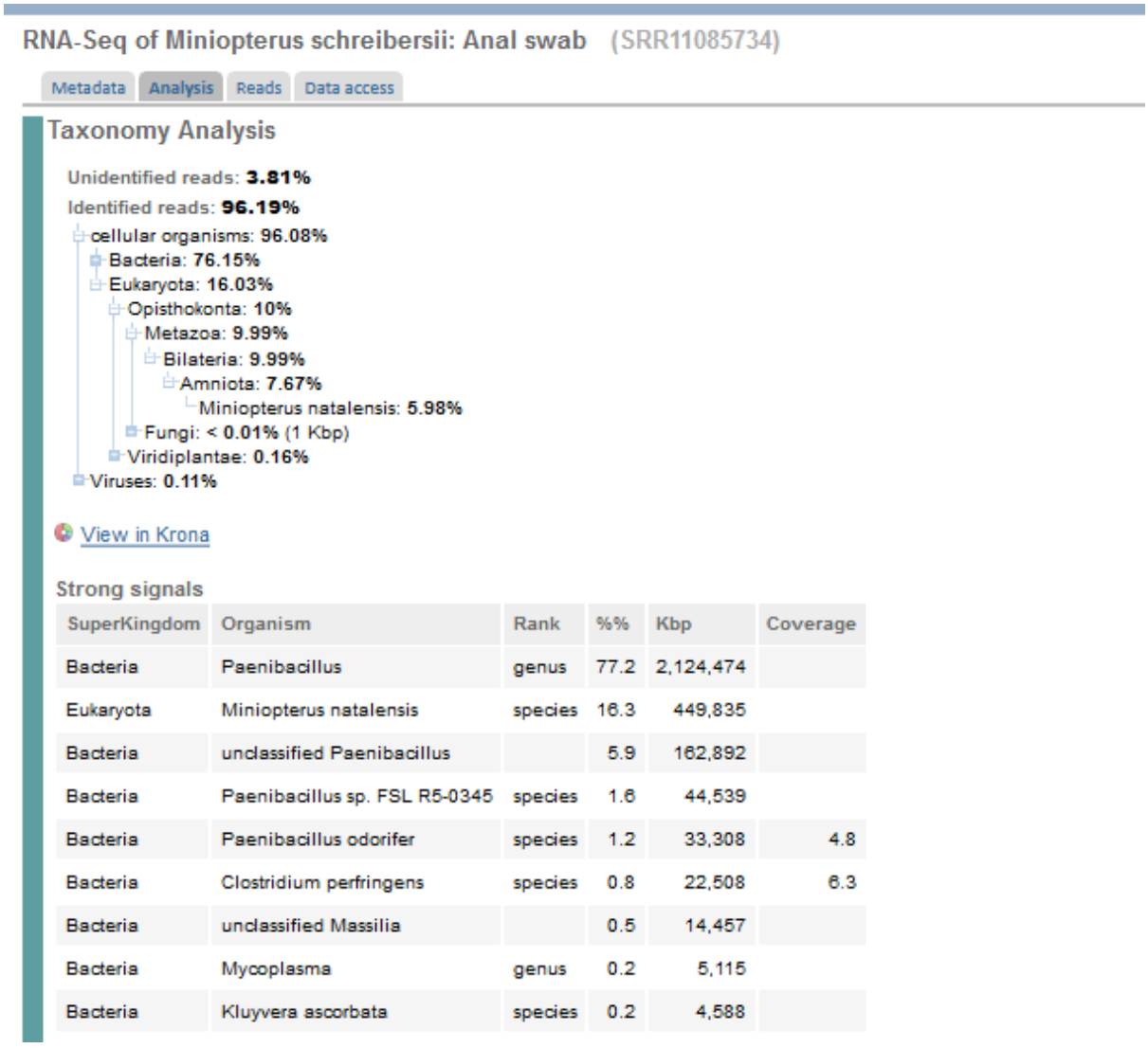
200

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

202

203

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



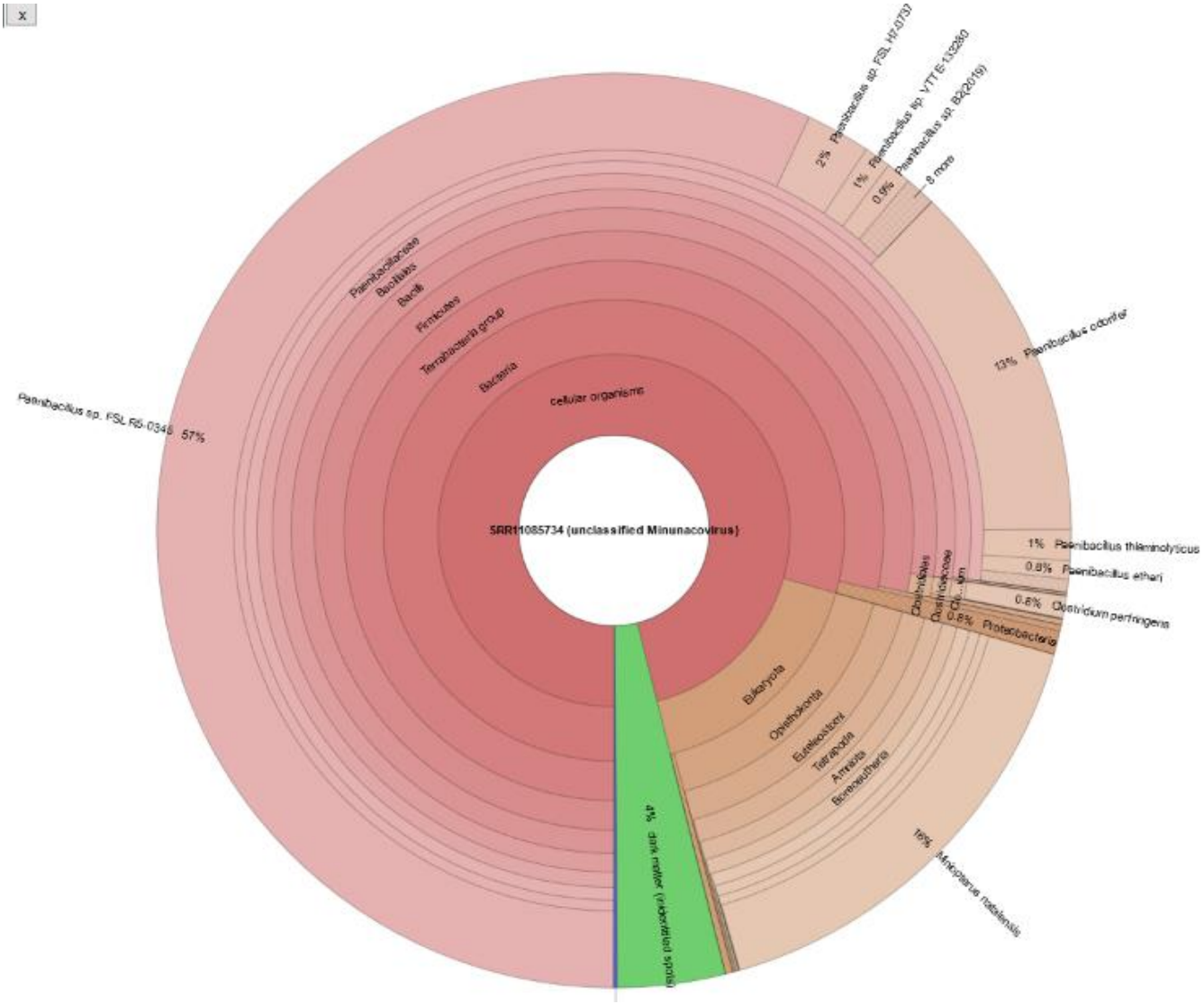
SuperKingdom	Organism	Rank	%%	Kbp	Coverage
Bacteria	Paenibacillus	genus	77.2	2,124,474	
Eukaryota	Miniopterus natalensis	species	16.3	449,835	
Bacteria	unclassified Paenibacillus		5.9	162,892	
Bacteria	Paenibacillus sp. FSL R5-0345	species	1.6	44,539	
Bacteria	Paenibacillus odorifer	species	1.2	33,308	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	

205

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

207

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



209
210

Fig. 4

Full ▾ Send to: ▾

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:

Arbuthnott, George, Calvert, Jonathan & Sherwell, Philip. 2020. https://www.thetimes.co.uk/article/seven-year-covid-trail-revealed-l5vxt7jqp. *The Sunday Times*.

Boni, M.F., Lemey, P & Jiang, X. et al. 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology,* https://doi.org/10.1038/s41564-020-0771-4.

Cohen, Jon 2020. Wuhan coronavirus hunter Shi Zhengli speaks out. *Science,* 369**,** 487-488.

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.

Wrobel, A.G., Benton, D.J. & Xu, P. et al. 2020. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.,* 27**,** 763-767.

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.