 doi:10.20944/preprints202008.0205.v3

Short Note

The anomalous nature of the fecal swab data, receptor binding domain and other questions in RaTG13 genome

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

Abstract:

RaTG13 (a bat derived SARS-like CoV) is the closest relative sequence of SARS-CoV-2 reported till date. The sample from which RaTG13 was sequenced was a bat fecal swab collected in 2013 from Tongguan, Mojiang, Yunnan province, China. The Illumina based sequence of RaTG13, MN996532.1, was deposited on 27th Jan 2020 and the raw data (Illumina), https://www.ncbi.nlm.nih.gov/sra/SRX7724752[accn]. There are discrepancies in dates about when the metagenome sequencing of RaTG13 sample was done (2018 or 2020), both stated by the same corresponding author. Comparison of the RNA Seq data of RaTG13 fecal swab to the corresponding data from the bat fecal swabs deposited by the same working group using the same methods indicated that the RaTG13 raw data seemed to be different in various aspects. The fecal swab sample showed abnormally less read of bacterial reads in the swab was exceptionally low, i.e. 0.7%, compared to the 20-90% abundance in other fecal swabs from bats processed by similar methods. Also, another raw data in the form of amplicon sequences was deposited in May 2020; however, the dates mentioned on the files of the sequenced amplicons were older (2017, 2018). The genome assembly of RaTG13 could not be done de-novo and the average coverage of the genome ~8%. Also, literature indicates that RaTG13 RBD cannot bind to *Rhinolophus* ACE-2 receptors. Collectively, the anomalies in the raw data of RaTG13 and other issues 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

COVID-19 has been a devastating pandemic affecting more than thirty three million people and killing more than one million people till date (30th September 2020). It has been reported that SARS-CoV2 is most similar to a bat derived coronavirus, recently introduced to the scientific community, named as RaTG13 (Zhou et al., 2020). The name RaTG13 has been introduced in 2020 along with SARS-CoV2. Dr. Zhengli Shi, the corresponding author for the same paper has clarified after almost 7 months of the publication, that this is synonymous to earlier collected sample and a SARS-like CoV called 4991 (Ge et al., 2016). As there is no live virus RaTG13 and no RNA sample available for the same, it is extremely important to verify if the sequence data from which the assembly was built show the necessary quality. Several studies have used the RaTG13 sequence for protein related experiments and other evolutionary analysis (Wrobel et al., 2020) (Boni et al., 2020), and many more upcoming papers are using the genome sequence.

According to the reference, the sequence of RaTG13 was retrieved by RNA sequencing using a next generation sequencing approach after it was found that a region in RdRp (370 bases) matched with a viral RdRp sequence derived from a *Rhinolophus affinis* fecal swab RNA collected in 2013 (Zhou et al., 2020). The RNA sample was that of a bat fecal swab collected in July 2013, from 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, Yunnan, China (Cohen, 2020). She also confirmed that the old name of RaTG13 virus according to the RdRp sequence of BtCoV/CoV4991, described earlier (Ge et al., 2016). In the same question and answers session she also clarified that the sample was sequenced using next generation sequencing in 2018. However, the sample is over after sequencing as per Dr. Zhengli Shi, the corresponding author (Cohen, 2020). Here, the same corresponding author stated two different years when RaTG13 metagenome was sequenced: 2018 (as per her most

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

recent statement (Cohen, 2020) and 2020, as per (Zhou et al., 2020)This discrepancy in the date should be noted and questions as to when exactly the metagenome was sequenced?

RaTG13 was first mentioned in 2020 (Zhou et al., 2020) and the full genome sequence was not available before 27th 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].

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) Rahalkar and Bahulikar 2020, accepted. 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).

Here are the basic discrepancies encountered after the analysis of the RaTG13 fecal swab data and other issues:

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 (Feb.13, 2020) is that the sample is recorded as being extracted from a BAL fluid (broncho alveolar lavage) (Fig. 1).

2. Metagenome analysis showed that a large part of the raw data showed low quality reads (47%), MG-RAST analysis. From the ~53% reads which passed the quality check, 44% of them were contributed by rRNA reads. As MG-RAST does not classify the eukaryotic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

sequences properly, we manually blasted the retrieved nucleotide sequences contributed by eukaryotes. Blast analysis of randomly chosen ~150 reads showed similarities to either predicted *in-silico* transcripts from a single sequence (*Rhinolophus ferrumequinum*, MPI-CBG mRhiFer1) NC_046302.1 genomic sequence or to a *Rhinolophus ferrumequinum* clone AC155226.4 (~40 kb clone) or other animals in some cases. No sequences showed similarity to *Rhinolophus affinis* sequences. Incidentally, the same group had deposited a *Rhinolophus affinis* anal swab SRA data, SRR11085736 (Figure 2), which would have some sequences from *Rhinolophus affinis*, however we found no sequences directly showing similarities to these reads or any other *Rhinolophus affinis*. Similar discrepancies in the raw data have been pointed out recently (Zhang, 2020). Also, the SRR11085736 data showed 91% bacterial reads, though the methods used for obtaining RaTG13 metagenome SRR11085797 and SRR11085736 are similar (NCBI records).

3. Another major discrepancy is that the RNA sequencing data shows extremely less abundance of bacteria, only 0.7% (according to the NCBI analysis) and similar value was found by our analysis in MG-RAST also. When we compared this to other fecal or anal swabs deposited by the same group, which used the same kits and same methods for RNA extraction and library generation, we found that the SRA data of all of these swabs showed the presence of at least 20-90% of bacterial reads. Bacteria are usually the highest constituents of gut flora and hence contribute to a high extent to a fecal sample.

4. The coronavirus sequence (RaTG13) contributed to ~0.003% of the total sequence reads. A total of 1762 raw reads were retrieved. However, we could not build a de-novo assembly from these reads but only when we used a reference sequence as the whole genome of RaTG13, we could build an assembly. There were less overlaps in a few regions, 2- 3 gaps and a coverage of ~8X. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

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.

5. No indications of amplicon sequencing have been given by Zhou et al 2020 or in any of the recent publications by the WIV workgroup. Amplicon sequencing files of RaTG13 (SRX8357956) seemed to have been submitted in May 2020, but the files have older dates from 2017 and 2018 (Figure 3).

6. There are two contrasting sequences for a single patch (spot 23 and spot 24), e.g. shows 95-96% 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. 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. 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 compared to the RdRp of 4991 or RaTG13 RdRp.

7. **RBD of RaTG13 genome does not bind to *Rhinolophus* ACE-2**

According to a recent paper, when RaTG13 receptor binding domain was checked for binding with various receptors, e.g. from bats, humans, pig and mouse, RBD, RaTG13 RBD sequence did not show binding efficiency to the tested bats (*R. macrotis* and *R. pusillus*). Instead it showed binding to mouse or rat RBD efficiently (Mou et al., 2020). This is particularly surprising as the virus was isolated from a bat fecal sample (in 2013). It has been noted that SARS-CoV-2 has a high nucleotide sequence identity with RaTG13-like virus except for the middle part of its genome encoding the spike protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

Conclusions:

RaTG13 beta coronavirus, which exists in the form of a genome sequence, is the closest relative of SARS-CoV-2 genome sequence 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 RdRp region of RaTG13, CoV4991 (KP8765496.1) was deposited in 2016, which seems to be much older than the genome data for RaTG13, MN996532.1, and 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]. RaTG13 sequence has been deposited after the COVID-19 outbreak and mentioned as sequenced in 2020, however, the corresponding author has recently told that it was sequenced in 2018. 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 and processed with the same methodology indicated that it is different from the other fecal/ anal swab raw data in several aspects. Metagenome analysis showed that a large part of the raw data showed low quality reads (47%). From the ~53% reads which passed the quality check, 44% of them were contributed by rRNA reads. Most of the retrieved protein sequences contributed by eukaryotes showed similarities to the predicted *in-silico* transcripts from a single sequence (*Rhinolophus ferrumequinum*, MPI-CBG mRhiFer1) NC_046302.1 genomic sequence or to a clone AC155226.4 as revealed by BLAST analysis, but not to *Rhinolophus affinis* sequences. The proportion of the bacterial reads in the swab was exceptionally low, i.e. 0.7%, which is abnormal, compared to the 20-90% bacterial abundance in other bat fecal swabs processed by the same methods (SRX). A complete de-NOVO assembly of RaTG13 could not be made from the 1762 retrieved viral reads which were of fairly good quality, even though the swab was 7 years old. A reference based assembly was done which indicated that some regions had

 doi:10.20944/preprints202008.0205.v3

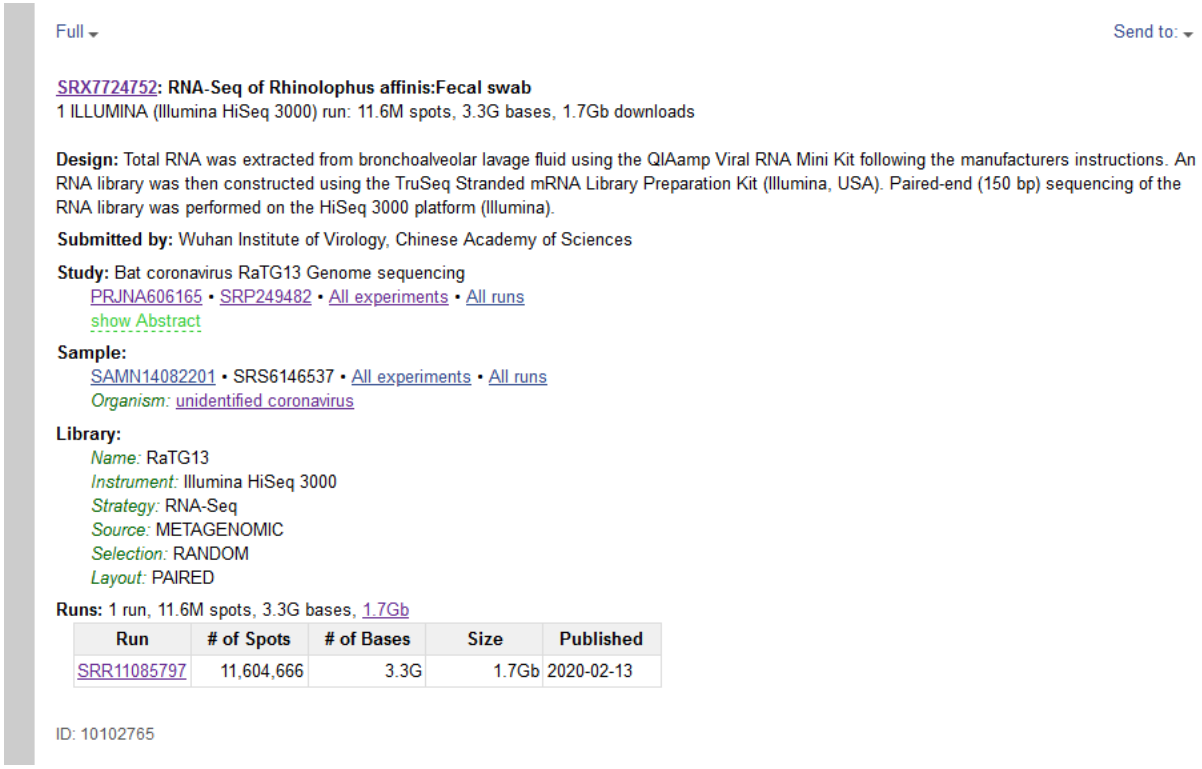
a single read coverage and the overall coverage was ~8X, quite low for a good assembly. Further, we also saw that recent studies have indicated that the RaTG13 genome shows a receptor binding domain which does not show binding to the *Rhinolophus* ACE-2. Another set of raw data associated with RaTG13, which seems to be amplicon sequencing of the genome (SRX8357956), submitted in May 2020 showed older dates (2017, 2018). Collectively, the anomalies in the raw data of RaTG13, the fact that the RBD of RaTG13 does not bind to bat receptors, and the date related confusion, pose an important question about the overall authenticity of the RaTG13 genome sequence.

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 and the genome content of RaTG13, we suggest that RaTG13 genome sequence should be interpreted with caution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

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

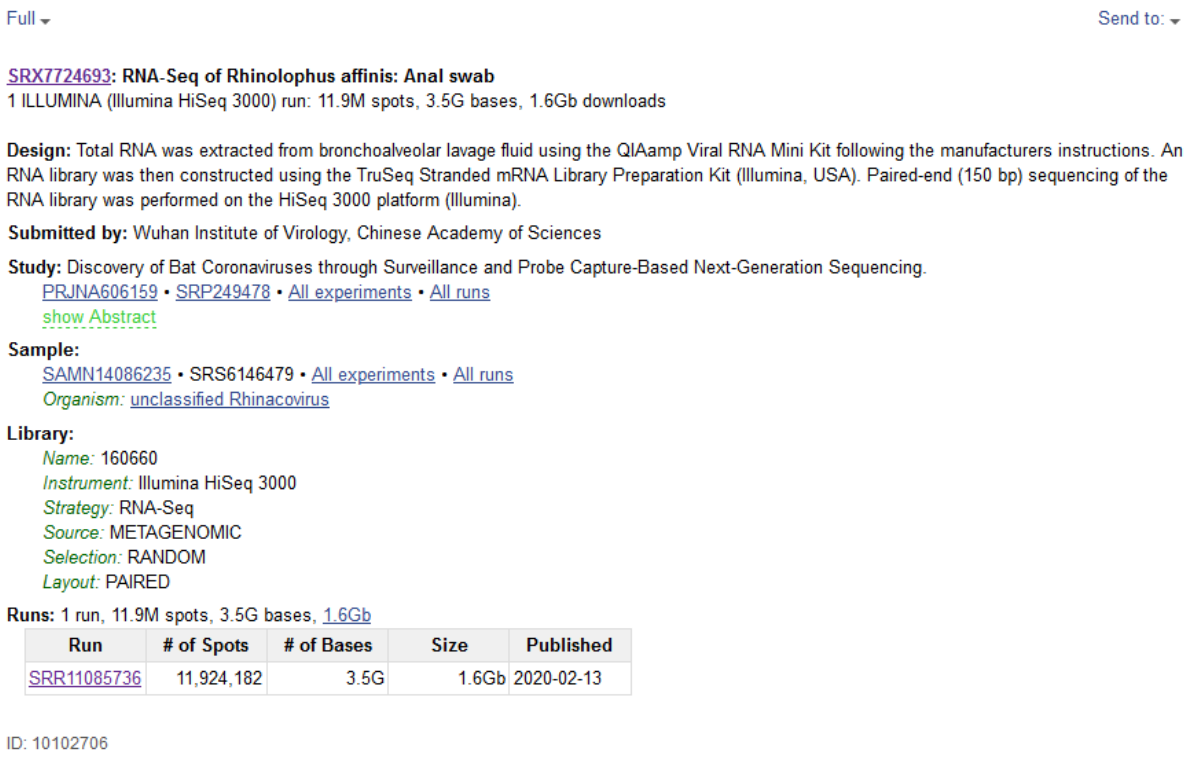
Runs: 1 run, 11.6M spots, 3.3G bases, 1.7Gb

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)**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3



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

179

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

181

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3

Taxonomy Analysis

Unidentified reads: **0.86%**

Identified reads: **99.14%**

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

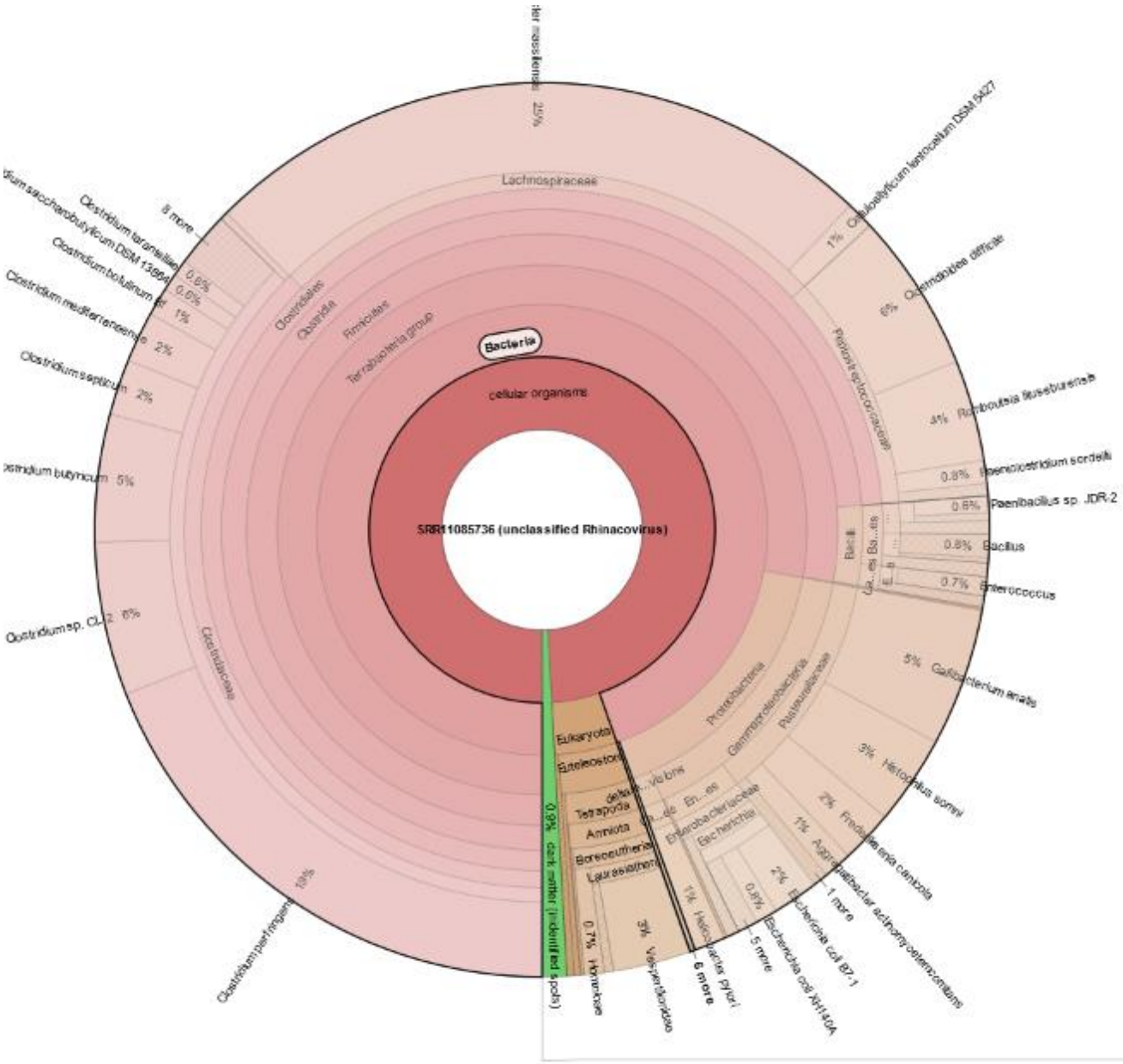
View in Krona

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202008.0205.v3



186

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

188

189

Fig. 3

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.

Mou, Huihui, Quinlan, Brian D., Haiyong Peng, Guo, Yan, Peng, Shoujiao, Zhang, Lizhou, Davis-Gardner, Meredith E., Gardner, Matthew R., Crynen, Gogce & Farzan, Michael 2020. Mutations from bat ACE2 orthologs markedly enhance ACE2-Fc neutralization of SARS-CoV-2. *bioRxiv*.

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.

 doi:10.20944/preprints202008.0205.v3

218 F. & Shi, Z. L. 2020. A pneumonia outbreak associated with a new coronavirus of probable
219 bat origin. *Nature,* 579**,** 270-273.

220