The seminal paper from the Wuhan Institute of Virology claiming SARS-CoV-2 probably originated in bats appears to contain a contrived specimen, an incomplete and inaccurate genomic assembly, and the signature of laboratory-derived synthetic biology

The coronavirus RaTG13 was purportedly identified in a bat "fecal" specimen that is probably not feces, has significant unresolved method-dependent genome sequence errors and an incomplete assembly with significant gaps, and has an anomalous base substitution pattern that has never been seen in nature but is routinely used in codon-optimized synthetic genome constructions performed in the laboratory

Author: Steven C. Quay, MD, PhD[1]

Abstract. The species of origin for the SARS-CoV-2 coronavirus that has caused the COVID-19 pandemic remains unknown after over six months of intense research by investigators around the world. The current consensus theory among the scientific community is that it originated in bats and transferred to humans either directly or through an intermediate species; no credible intermediate species exists at this time. The suggested origin early on from a Wuhan "wet market" has been determined to be a red herring and the pangolin is no longer considered a likely intermediate by the virology community.

The basis for the hypothesis that SARS-CoV-2 probably evolved from bats initially came from a February 2020 paper[2] from Dr. Zheng-Li Shi's laboratory at the Wuhan Institute of Virology (WIV). In that paper the Wuhan laboratory made two claims: 1), "a bat fecal sample collected from Tongguan town, Mojiang county in Yunnan province in 2013" contained a coronavirus, originally designated "Rhinolophus bat coronavirus BtCoV/4991[3]" in 2016 but renamed in their paper, RaTG13; and 2), the genomes of RaTG13 and SARS-CoV-2 had an overall identity of 96.2%, making it the closest match to SARS-CoV-2 of any coronavirus identified at that time. RaTG13 remains the closest match to SARS-CoV-2 at the current time.

In this paper I document that:

1) The RaTG13 specimen was not a bat fecal specimen, based on a comparison of the relative bacterial and eukaryotic genetic material in the purported fecal specimen to nine authentic bat fecal specimens collected in the same field visits as RaTG13 was collected by the Wuhan laboratory, run on the same Illumina instrument (id ST-J00123), and published in a second paper in February 2020.[15] While the authentic bat fecal samples

[1]Email: **Steven@DrQuay.com;** 107 Spring Street, Seattle, WA 98104. ORCID: 0000-0002-0363-7651

[2] 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 .

[3] A Coronavirus BtCoV/4991 Genbank entry by Dr. Shi records: organism="Rhinolophus bat coronavirus BtCoV/4991." In July 2020 she wrote: "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

were, as expected, largely bacterial (specifically, 65% bacteria and 12% eukaryotic genetic sequences), the purported RaTG13 specimen had a reversed composition, with mostly eukaryotic genes and almost no bacterial genetic material (0.7% bacteria and 68% eukaryotic). The RaTG13 specimen was also only 0.01% virus genes compared to an average of 1.4% for authentic bat fecal specimens. A Krona analysis identified 3% primate sequences consistent with VERO cell contamination, the standard monkey cell culture used for coronavirus research, including at the Wuhan laboratory. Based on using the mean and standard deviation of the nine authentic bat fecal specimens from the Wuhan laboratory, the probability that RaTG13 came from a true fecal sample but had the composition reported by the Wuhan laboratory is one in thirteen million;

2) According to multiple references, RaTG13 was identified via Sanger dideoxy sequencing before 2016, partially sequenced by amplicon sequencing in 2017 and 2018, and then complete sequencing and assembly by RNA-Seq in 2020, although some reports from WIV suggest the timing of the RNA-Seq experiments may have been performed earlier than 2020. In any case, a Blast analysis of sequences from the amplicon and RNA-Seq experiments indicates an approximate 5% nucleotide difference, 50-fold higher than the technical error rate for RNA-Seq of about 0.1%. At least two gaps of over 60 base-pairs, with no coverage in the RNA-Seq data, were easily identified. The incomplete assembly and anomalous, method-dependent sequence divergence for RaTG13 is troublesome;

3) The pattern of synonymous to non-synonymous (S/NS) sequence differences between RaTG13 and SARS-CoV-2 in a 2201 nucleotide region flanking the S1/S2 junction of the Spike Protein records 112 synonymous mutation differences with only three non-synonymous changes. Based on the S/NS mutational frequencies elsewhere in these two genomes and generally in other coronaviruses the probability that this mutation pattern arose naturally is approximately one in ten million. A similar pattern of unnatural S/SN substitutions was seen in a 10,818 nt region of the pp1ab gene. This pp1ab gene pattern has a probability of occurring naturally of less than one in 100 billion. A total of four regions of the RaTG13 genome, coding for 7,938 nt and about one-quarter of the entire genome, contain over 200 synonymous mutations without a single non-synonymous mutation. This has a probability of one in 10^{-17}. A possible explanation, the absolute criticality of the specific amino acid sequence in the regions which might make a non-synonymous change non-infective, is ruled out by the rapid appearance of an abundance of non-synonymous mutations in these very regions when examining the over 80,000 human SARS-CoV-2 specimens sequenced to date. An alternative hypothesis, that this arose by codon substitution is examined. It is demonstrated, by example from a published codon-optimized SARS-Cov-2 Spike Protein experiment, that the anomalous S/SN pattern is precisely the pattern which is produced, by design, when synthetic biology is used and represents a signature of laboratory construction.

Based on the findings concerning the RaTG13 data, including anomalies and inconsistent statements about RaTG13, its origin, renaming, and sequencing timing; the finding that the specimen it is purported to have come from is not bat feces and has a signature of cell culture contamination; the unexplained method-dependent 5% sequence difference for RaTG13; and the S/SN mutation pattern reported, which to my knowledge has never been seen in nature, it can be concluded that RaTG13 is not a pristine biological entity but shows evidence of genetic manipulation in the laboratory.

Until a satisfactory explanation of the findings in this paper have been offered by the Wuhan laboratory, all hypotheses of the proximal origin of the entry of SARS-CoV-2 into the human population should now include the likelihood that the seminal paper contains contrived data. For example, the hypothesis that SARS-CoV-2 was the subject of laboratory research and at some point escaped the laboratory should be included in the narrative of the origin of SARS-CoV-2 research.

Introduction. Since the first reported patient on December 1, 2019 with a SARS-CoV-2 infection, the virus has caused a pandemic that has led to twenty-five million cases worldwide and over 840,000 deaths as of August 30, 2020. To make progress on treating this disease and preventing the next viral outbreak, knowing the origin of the virus and how it entered the human population is critical.

On February 3, 2020 a paper was published from the Wuhan Institute of Virology that identified a bat coronavirus, RaTG13, as having a 96.2% identity to SARS-CoV-2, quickly providing support for a zoonotic origin, either from bats directly or from bats to humans through an unknown intermediary species. If true, this would replicate the model of SARS-CoV 2003 in which the transmission was from bats to civets to humans and for MERS in which the transmission was from bats to camels to humans. At the time of this paper and through August 30, 2020, no other virus has been identified with a closer sequence homology to SARS-CoV-2 than RaTG13. The publication containing the RaTG13 sequence has been cited over 1600 times in the six months since publication. None of these studies contain research on the isolated virus itself since the virus has never been isolated or cultured. It was apparently found in only one sample from 2013 and that sample has been exhausted.[4]

An examination of the raw data associated with RaTG13 immediately identified serious anomalies, bringing into question the existence of RaTG13 as a biological entity of completely nature origin.

[4] Dr. Shi Science interview July 2020

Materials and Methods.

GenBank accession URL table for sequences used in this paper.

The GenBank accession URLs for the specimens, raw reads, and sequences that are used in this paper are contained in the following Table, which can be used to reach the raw data.

Descriptor	URL Hyperlink
SARS-CoV-2 reference sequence in GenBank	SARS-CoV-2 complete genome
Bat coronavirus RaTG13, complete genome, Genbank	RaTG13 complete genome
RaTG13 purported bat fecal specimen	SRR11085797
Rhinolophus bat coronavirus BtCoV/4991 RNA-dependent RNA polymerase (RdRp) gene, partial cds	BtCoV/4991 RdRp gene
SRX8357956: amplicon_sequences of RaTG13	Specimen descriptor
RNA-Seq data for RaTG13	RNA-Seq data for RaTG13
Reference fecal bat specimens from WIV	SRR11085736
Reference fecal bat specimens from WIV	SRR11085734
Reference fecal bat specimens from WIV	SRR11085737
Reference fecal bat specimens from WIV	SRR11085733
Reference fecal bat specimens from WIV	SRR11085735
Reference fecal bat specimens from WIV	SRR11085738
Reference fecal bat specimens from WIV	SRR11085739
Reference fecal bat specimens from WIV	SRR11085740
Reference fecal bat specimens from WIV	SRR11085741

Below is a screen shot of the GenBank entry for the purported specimen from which RaTG13 was identified and upon which RNA-Seq was performed. While the title claims it is a “Rhinolophus affinis fecal swab” specimen it also records in the design of work entry that “(t)otal RNA was extracted from bronchoalveolar lavage fluid.” These descriptions are clearly inconsistent.

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

Apparent missing amplicon reads for RaTG13 in GenBank.

There are 33 amplicon reads in GenBank for RaTG13 from experiments recorded as having been performed in 2017 and 2018. A file naming pattern was noticed among the data sets which suggests there may be amplicon runs that were not deposited in GenBank. These files, if related to RaTG13, may contain useful sequence data and an effort should be made to retrieve them and, if appropriate, upload them to GenBank. A Table with the apparently missing data (yellow) is shown here.

Date	Amplicon file name endings							
3-Jun-17	A07	A08						
17-Jun-17	A05	A06						
20-Jun-17						F03	G03	H03
27-Sep-18	A06	B06	C06		E05	F05	G05/G06	H05/H06
29-Sep-18				D05	E05		G04	H04
30-Sep-18	A02	B11						
8-Oct-18			C11				G10	H11
11-Oct-18	A12	B12						
14-Oct-18	A02	B02	C02	D02				

Relationship of *Rhinolophus* bat coronavirus BtCoV/4991 and Bat coronavirus RaTG13.

The Wuhan laboratory has reported on the bat coronaviruses, BtCoV/4991 and RaTG13, in two peer-reviewed publications, one in 2016 and one in February 2020.[5] They have submitted three entries to GenBank for these two viruses, in 2016, February 2020, and May 2020.[6] The GenBank entries confirm sequencing experiments using Sanger dideoxy sequencing in 2016, PCR-generated amplicon sequencing performed on an AB 310 Genetic Analyzer in 2017 and 2018, and RNA-seq performed on an Illumina HiSeq 3000 (instrument id ST-J00123) in 2020. A single GISAID entry records that the RNA-seq data was obtained from an original specimen without passage.[7] This is an important detail since evidence of primate sequences, consistent with VERO cell contamination, is found in this specimen, as reported below, which would suggest laboratory passage.

None of these disclosures report that BtCoV/4991 and RaTG13 are the same coronavirus, simply renamed. This information was only disclosed in a written Question and Answer publication from *Science* magazine by Dr. Shi on July 31, 2020.[4, 8] Given this disclosure months after the original

[5] 2016 Virologica Sinica paper and February 2020 Nature paper

[6] RaTG13 complete genome Feb 2020, Raw sequence reads for RaTG13 published Feb 2020, Amplicon reads for RaTG13 from 2017 and 2018 published in May 2020.

[7] The GISAID entry is EPI_ISL_402131.

[8] Dr. Shi wrote: "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

publication concerning RaTG13 in *Nature* it is possible that the omission of the original publication and sequence data concerning BtCoV/4991 violated the "Reporting standards and availability of data, materials, code and protocols" required for *Nature* publications.[9]

The February 2020 papers uses the RNA-Seq data for RaTG13 genome determination but fails to disclose the previous data obtained by Sanger dideoxy sequencing in 2016 and by amplicon sequencing in 2017 and 2018. Since these unrecorded data establish method-dependent sequencing differences of up to 4% the failure to disclose this data or to reconcile these differences is troubling.

In addition, the raw assembly accession data for RaTG13 are not described or linked to the Genbank entry, MN669532, and also no assembly method is specified in the raw data SRX7724752 12 and the Illumina run. And the amplicon sequencing data has sequence gaps of approximately 20% of the genome. Therefore, no primary assembly data has been made available by the WIV for the RaTG13 genome. This is contrary to the *Nature* Reporting Standards[9] as they state: "When publishing reference genomes, the assembly must be made available in addition to the sequence reads."

Relationship of RaTG13 and SARS-CoV-2.

There have been two descriptions of the process by which the RaTG13 genome was identified as closely homologous to SARS-CoV-2. These seem to be inconsistent with each other.

In the February 2020 *Nature* paper[5] it states:

"We then found that a short region of RNA-dependent RNA polymerase (RdRp) from a bat coronavirus (BatCoV RaTG13)—which was previously detected in Rhinolophus affinis from Yunnan province—showed high sequence identity to 2019-nCoV. We carried out full-length sequencing on this RNA sample (GISAID accession number EPI_ISL_402131). Simplot analysis showed that 2019-nCoV was highly similar throughout the genome to RaTG13, with an overall genome sequence identity of 96.2%."

In a July 2020 interview the process was described:

"We detected the virus by pan-coronavirus RT-PCR in a bat fecal sample collected from Tongguan town, Mojiang county in Yunnan province in 2013, and obtained its partial RdRp sequence. Because the low similarity of this virus to SARS-CoV, we did not pay special attention to this sequence. In 2018, as the NGS sequencing technology and capability in our lab was improved, we did further sequencing of the virus using our remaining samples, and obtained the full-length genome sequence of RaTG13 except the 15 nucleotides at the 5' end. As the sample was used many times for the purpose of viral nucleic acid extraction, there was no more sample after we finished genome sequencing, and we did not do virus isolation and other studies on it. Among all the bat samples we collected, the RaTG13 virus was detected in only one single sample. In 2020,

[9] Nature research reporting standards for availability of data

we compared the sequence of SARS-CoV-2 and our unpublished bat coronavirus sequences and found it shared a 96.2% identity with RaTG13. RaTG13 has never been isolated or cultured."

If the full-length genome of RaTG13 was available by 2018 it is unclear why a database search within the WIV for coronaviruses that resembled SARS-CoV-2 would lead to identifying the 370-nt segment representing the RdRp gene (as stated in the February paper) but not the full length RaTG13 genome (which was stated to have been sequenced by 2018). In addition, an assembly of all available amplicon data for RaTG13 from 2017 and 2018 contains gaps of approximately 20% of the genome. If the sample was completely consumed during the 2017-8 sequencing it is unclear how RNA-Seq was conducted in 2020 to permit the full-length genome to be determined.

Analytical methods. Taxonomy of specimens was determined in the NCBI Sequence Read Archive and KRONA.[10] Blast was used for sequence alignment and comparisons.[11]

To evaluate the data from the bat species relative to the RaTG13 fecal sample analysis, the latter was treated as a fixed result with the comparison to the taxonomy results of the nine bat feces specimens. It also was noted that the data were clearly right skewed (and descriptively both mean/median and standard deviation/interquartile range were used). Therefore, a non-parametric procedure, the Wilcoxon signed-rank test was used with the p-value calculated by an exact procedure because of the small sample size. Considering the synonymous to non-synonymous mutation frequency and how to evaluate that for the various protein coding regions of the virus, it was noted that for all of the genes pooled, the ratio of the synonymous to non-synonymous regions was approximately 0.83. To analyze the corresponding distribution for each gene, we assumed that each mutation was an independent observation from a Bernoulli random variable and, therefore the number of synonymous mutations in the gene would have a binomial distribution (with probability 0.83). A probability was then computed for the actual number of synonymous mutations on this basis (the probability was determined on a one-sided basis, i.e. excess mutations, and was calculated as a strict inequality).

Results.

Original characterization of RaBtCoV/4991 (RaTG13) and related bat fecal specimen.

In 2016 Dr. Shi and colleagues published a paper entitled, "Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft[12]" in which a number of novel bat coronaviruses were isolated from bat fecal specimens collected during 2012 and 2013. The viruses were named, according to the paper, in the following fashion:

> "The positive samples detected in this study were named using the abbreviated bat species name plus the bat sample number abbreviation. For example, a virus detected

[10] NCBI Sequence Archive
[11] Blast alignment
[12] Xing-Yi Ge, et. al., Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft, Virologica Sinica, 2016, 31 (1): 31–40. DOI: 10.1007/s12250-016-3713-9

> from *Rhinolophus sinicus* in sample number 4017 was named RsBtCoV/4017. If the bat was co-infected by two different coronaviruses, numbers were appended to the sample names, such as RsBtCoV/4017-1 and RsBtCoV/4017-2."

In the July 2020 interview Dr. Shi wrote:

> "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

The 2016 and 2020 statements about the naming of virus RsBtCoV/4991 appear inconsistent with each other.

Of the 152 coronaviruses identified, 150 were classified as alphacoronaviruses while only two were classified as betacoronaviruses, HiBtCoV/3740-2 and RaBtCoV/4991. The naming convention from the paper means this latter coronavirus was identified in a fecal specimen from a *Rhinolophus affinis* bat and was sample number 4991.

The latter virus was described in the paper as follows:

> "Virus RaBtCoV/4991 was detected in a R. affinis sample and was related to SL-CoV. The conserved 440-bp RdRp fragment of RaBtCoV/4991 had 89% nt identity and 95% aa identity with SL-CoV Rs672. In the phylogenetic tree, RaBtCoV/4991 showed more divergence from human SARS-CoV than other bat SL-CoVs and could be considered as a new strain of this virus lineage."

The Genbank accession number for RaBtCoV/4991 is MN_KP876546.1 and in Genbank it is identified as having been collected in July 2013 as a "feces/swabs" specimen.

The RATG13 genome sequence was assembled from low coverage RNA-Seq data.

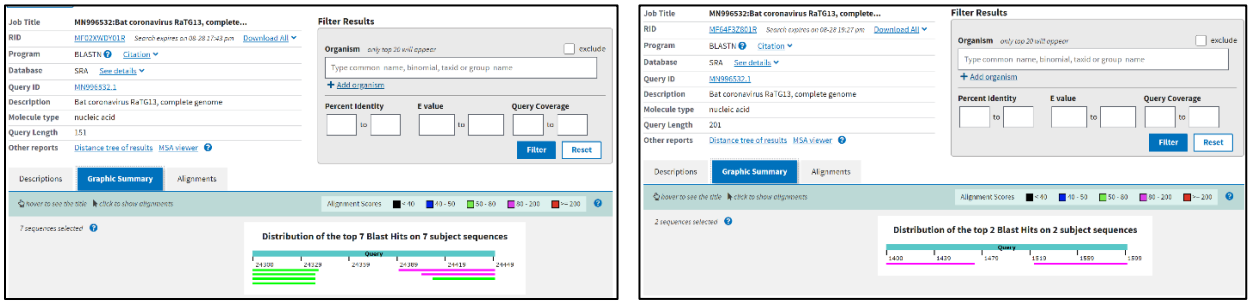
A Blast analysis of the RaTG13 genome against SRR11085797 retrieved about 1700 reads which covers only about 252,000 nt of the total reads of 3.3 Gb. Since the genome size of RaTG13 is known to be about 30,000 nt this represents an 8-fold coverage, typically insufficient for a definitive assembly. For example, some have suggested a 30-fold coverage is necessary to create high quality assemblies.[13]

At an eight-fold coverage and based on the typical practice of having four or more reads to call a SNP,[14] the 8-fold coverage of RaTG13 would have 4.2% bases or about 1260 calls of less than 4 reads and about 10 bases would be missed completely, with no calls at all.

[13] Sims, D. *et al.* Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews – Genetics. (2014) 15: 121-132. doi:10.1038/nrg3642.
[14] Illumina Technical Bulletin Call Coverage

A Blast of the RaTG13 published genome onto the RNA-Seq data documents at least two 60 base-pair gaps with no coverage, precluding a complete assembly.

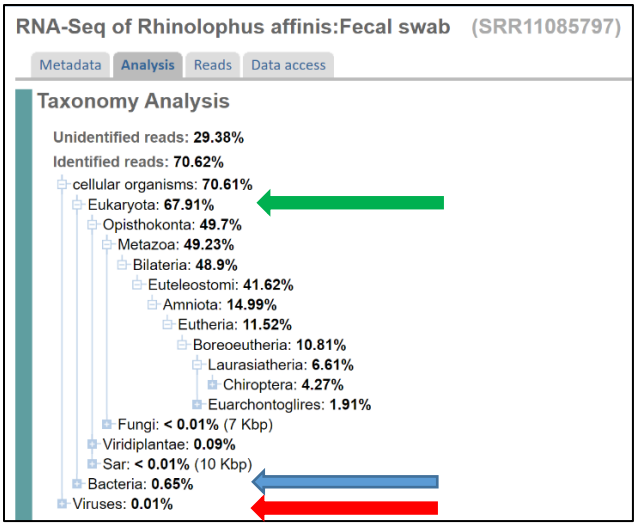
Given the low coverage in the RNA-Seq data, an exploratory, non-exhaustive Blast search was conducted against the published RaTG13 sequence. Two gaps of over 60 nt, shown below, were easily found:



It is conceivable there are additional gaps but the above two are sufficient to document that the complete RaTG13 genome sequence could not have been assembled solely from the RNA-Seq data, as stated.[2]

Taxonomy analysis of the RaTG specimen is inconsistent with being from bat feces and shows evidence of laboratory cell culture contamination.

According to the Wuhan laboratory, the RaTG13 coronavirus was a fecal swab specimen collected from a *Rhinolophus affinis* bat in 2013. Unexpectedly, (Text-Figure below) the taxonomy analysis is primarily eukaryotic (green arrow; 67.91%) with only traces of bacteria (blue arrow; 0.65%). The viral genomes also make only a trace contribution (red arrow; 0.01%):



Taxonomy analysis for RaTG13 data SRR11085797

To compare this specimen composition to bat fecal specimens collected by Dr. Shi and her WIV colleagues and analyzed in other studies, a paper from Dr. Shi's laboratory, also published in

February 2020, was identified. In this paper, entitled, "Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing,"[15] a total of nine specimens "collected during previous bat CoV surveillance projects, (were) extracted from bat rectal swabs." According to the Methods section in this paper, the "previous bat CoV surveillance projects" include the field work in 2013 when the RaTG13 was said to have been collected. The comparison below is thus the same specimens collected on the same field surveillance projects by the same investigators from the Wuhan laboratory and sequenced on the same Illumina instrument. These nine specimens will be referred to as "reference fecal specimens" henceforth.

The following Text-Table compares the taxonomical analysis of the RaTG13 and reference fecal specimens. The reference fecal specimens have an average eukaryotic genome content of about 12% while RaTG's eukaryotic content was 68%. On the other hand, the most abundant genes in the reference fecal specimens were bacterial, with an average of 65%; RaTG13 had less than 1% bacterial genes. And finally, the reference fecal specimens had 1.57% virus genes compared to the 0.01% virus genes of RaTG13.

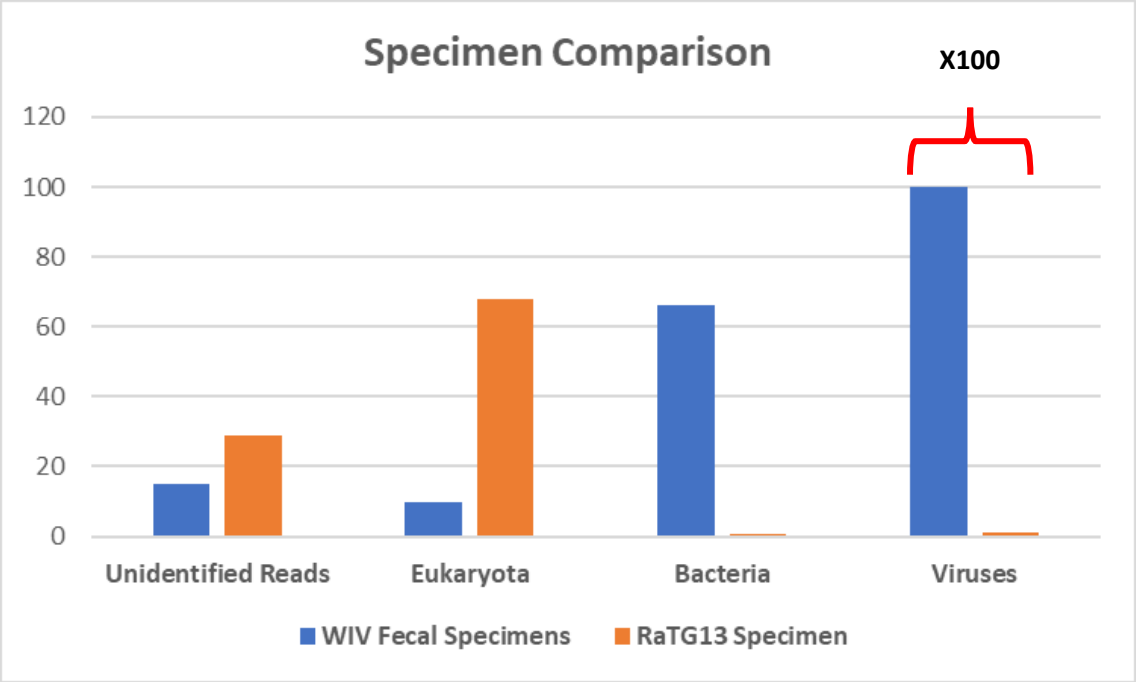
Specimen ID	Specimen Type	Unidentified Reads	Eukaryota	Bacteria	Viruses	Sum
SRR11085736	*Rhinolophus affinis*	0.86	4.36	91.07	0.03	96.32
SRR11085734	*Miniopterus schreibersii*	3.81	16.03	76.15	0.11	96.1
SRR11085737	*Scotophilus kuhlii*	17.98	8.59	67.81	2.19	96.6
SRR11085733	*Hipposideros larvatus*	13.27	27.99	42.96	4.1	88.32
SRR11085735	*Hipposideros pomona*	34.33	7.96	54.78	0.71	97.78
SRR11085738	*Pipistrellus abramus*	20.33	21.44	47.3	6.45	95.52
SRR11085739	*Tylonycteris pachypus*	61.75	14.34	20.06	0.06	96.21
SRR11085740	*Miniopterus pusillus*	0.78	1.46	99.22	0.05	101.51
SRR11085741	*Rousettus aegyptiacus*	6.44	2.59	88.36	0.45	97.84
Mean +/- SD	Nine bat feces specimens	17.73+/-19.79	11.64+/-9.02	65.30+/-26.10	1.57+/-2.28	96.24+/-3.45
Median +/- IQR	Nine bat feces specimens	13.27+/-24.995	8.59+/-15.26	67.81+/-41.58	0.45+/-3.09	96.32+/-2.00
SRR11085797	**RaTG13 fecal specimen**	**29.38**	**67.91**	**0.65**	**0.01**	**97.95**
	P-value (exact Wilcoxon signed-rank test)	0.16	**0.0039**	**0.0048**	**0.0039**	0.098

As shown in the Text-Table above the RaTG13 specimen is significantly different from the reference fecal specimens in composition. The probabilities for each category, eukaryote, bacteria, and virus, are individually highly statistically significant. They are also independent of each other and therefore the overall probability that RaTG13 has the composition of eukaryote, bacteria, and virus genes that was reported by the Wuhan laboratory but is actually from an authentic bat fecal specimen is less than one in 13 million.

The alternative conclusion is that this sample was not a fecal specimen but was contrived. The data cannot, however, distinguish between a non-fecal specimen that came from true field work on the one hand and a specimen created *de novo* in the laboratory on the other hand.

[15] Discovery of bat coronaviruses through surveillance and probe capture-based next-generation sequencing

A graphical comparison of the above data is shown below and visually shows the significant differences between the WIV fecal specimens and the RaTG13 specimen, despite the claim they were collected in the same field surveillance trips:



Another comparison can be made between the reference fecal specimens and the RaTG13 specimen by looking at the taxonomy of the nine to twelve "strong signals" identified on the NCBI Sequence Read Archive. The following Text-Table is a summary of these findings.

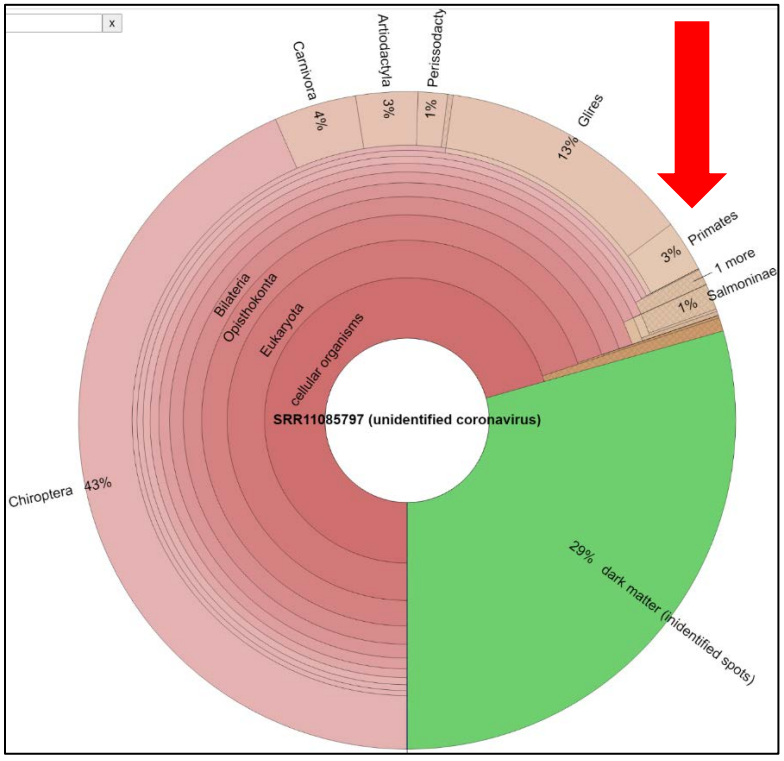
Specimen	The identity of the Strong Signals in the Specimens		
	Bacteria	Eukaryotes	Viruses
Rhinolophus affinis anal swab (SRR11085736)	92%	One magnaorder of placental mammals, includes bat	None
Miniopterus schreibersii anal swab (SRR11085734)	88%	**One bat**, the host bat, Miniopterus sp.	None
Scotophilus kuhlii anal swab (SRR11085737)	56%	**Two bats**, mouse-eared and big brown bats.	Two viruses, kobuvirus (host includes bats) and a Scotophilus kuhlii coronavirus
Hipposideros larvatus anal swab (SRR11085733)	56%	**One bat,** the host bat, Hipposideros sp. and one rodent.	Hipposideros pomona bat coronavirus
Hipposideros pomona: Anal swab (SRR11085735)	78%	**One bat**, the host bat, Hipposideros sp.	None
Pipistrellus abramus: Anal swab (SRR11085738)	73%	**Two bats**, the big brown bat and the mouse-eared bat.	Pipistrellus abramus bat coronavirus
Tylonycteris pachypus: Anal swab (SRR11085739)	67%	**Three bat**s, the microbat, the great roundleaf bat, and a superorder of mammals, which includes bats.	None
Miniopterus pusillus: Anal swab (SRR11085740)	89%	**One bat**, the Natal long-fingered bat.	None
Rousettus aegyptiacus: Anal swab (SRR11085741)	91%	One magnaorder of placental mammals, includes bats.	None
Average	77%		
RaTG13 Rhinolophus affinis:Fecal swab (SRR11085797)	None	All nine strong signals are eukaryotes. **Five bats**, the Great Roundleaf bat, resident of China, the Egyptian fruit bat, which is not found in China, a megabat, mouse-eared bat, and bent-winged bat. Two marmots, the Alpine marmot from Europe and the Yellow-bellied marmot of North America.The paraorder of whales. The red fox.	None

As can be seen, while the strong signals in the authentic specimens contain 56% to 92% (average 77%) bacterial signals, the RaTG13 specimen has no bacteria among the nine strong signals. Most specimens do not have virus strong signals but the three that do are host-related coronaviruses (four) or one host-related kobuvirus.

RaTG13 has no viral strong signals. Among the reference specimens with eukaryotic strong signals, they are either bat-related genes (eleven) or higher order taxonomy signals that include bats (three). There is one anomalous rodent-related signal among the reference specimens.

The RaTG13 specimen is again an outlier with all nine strong signals arising from eukaryotic genes. Five of the nine signals are bats, some resident to China and some with non-Chinese host ranges. Surprisingly, unlike three of the reference bat signals which are identified as host-related, the RaTG13 specimen did not contain *Rhinolophus* sp. host-related strong signals. The remaining four strong signals are marmot-related genes (two), whale-related gene (one), and red fox-related gene (one).

Finally, a Krona analysis (below) identifies 3% primate sequences (red arrow) in the RaTG13 sequence data. This is consistent with contamination by the standard laboratory coronavirus cell culture system, the VERO monkey kidney cell line.



Source: Krona analysis of RaTG13 specimen

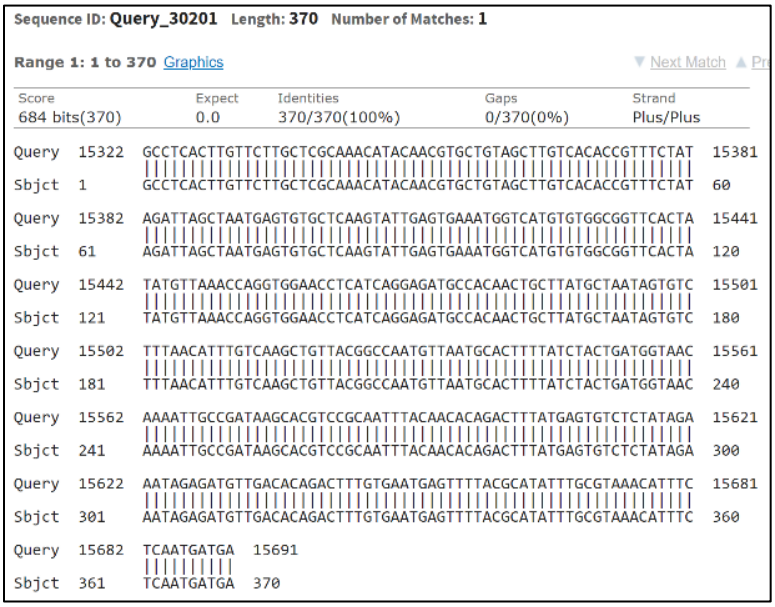
It is unclear why these obviously anomalous findings were not detected during the peer-review process prior to publication of this important work. At this point, an explanation is needed from the WIV to refute the conclusion that the specimen identified as the source of RaTG13 is **not** a

bat fecal/anal specimen and that the primate genetic material is consistent with a VERO cell contaminated specimen.

Method-related nt base substitutions in RaTG13.

The original Sanger dideoxy RdRp sequence reported in 2016 is homologous to RNA-seq data from 2020 but is non-homologous to amplicon sequencing data from 2017 and 2018.

As expected, a comparison of the 2016 RdRp GenBank sequence for BtCoV/4991 obtained by Sanger dideoxy sequencing with the RNA-seq sequencing of RaTG13 reported in *Nature* shows 100% identity over the 370 nt segment.



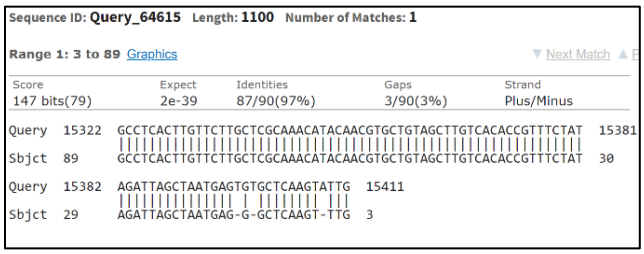
Sequence ID: **Query_30201** Length: **370** Number of Matches: **1**

Range 1: 1 to 370 Graphics ▼ Next Match ▲ Pr

Score	Expect	Identities	Gaps	Strand
684 bits(370)	0.0	370/370(100%)	0/370(0%)	Plus/Plus

```
Query  15322  GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  15381
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1      GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  60

Query  15382  AGATTAGCTAATGAGTGTGCTCAAGTATTGAGTGAAATGGTCATGTGTGGCGGTTCACTA  15441
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61     AGATTAGCTAATGAGTGTGCTCAAGTATTGAGTGAAATGGTCATGTGTGGCGGTTCACTA  120

Query  15442  TATGTTAAACCAGGTGGAACCTCATCAGGAGATGCCACAACTGCTTATGCTAATAGTGTC  15501
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121    TATGTTAAACCAGGTGGAACCTCATCAGGAGATGCCACAACTGCTTATGCTAATAGTGTC  180

Query  15502  TTTAACATTTGTCAAGCTGTTACGGCCAATGTTAATGCACTTTTATCTACTGATGGTAAC  15561
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181    TTTAACATTTGTCAAGCTGTTACGGCCAATGTTAATGCACTTTTATCTACTGATGGTAAC  240

Query  15562  AAAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGA  15621
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241    AAAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGA  300

Query  15622  AATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  15681
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301    AATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  360

Query  15682  TCAATGATGA  15691
              ||||||||||
Sbjct  361    TCAATGATGA  370
```

Surprisingly, the two amplicon sequences from 2017 that partially cover the 370 nt RdRp region have four base substitutions or gaps over a total segment of 219 nt (2% divergence).

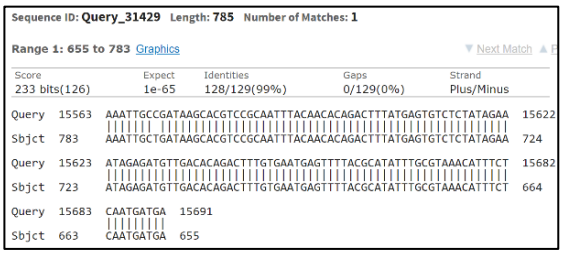


Sequence ID: **Query_64615** Length: **1100** Number of Matches: **1**

Range 1: 3 to 89 Graphics ▼ Next Match ▲ P

Score	Expect	Identities	Gaps	Strand
147 bits(79)	2e-39	87/90(97%)	3/90(3%)	Plus/Minus

```
Query  15322  GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  15381
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  89     GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  30

Query  15382  AGATTAGCTAATGAGTGTGCTCAAGTATTG  15411
              |||||||||||||| | |||||||| |||
Sbjct  29     AGATTAGCTAATGAG-G-GCTCAAGT-TTG  3
```



Sequence ID: **Query_31429** Length: **785** Number of Matches: **1**

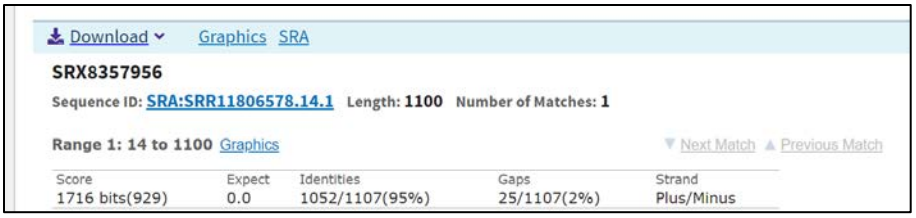
Range 1: 655 to 783 Graphics ▼ Next Match ▲ P

Score	Expect	Identities	Gaps	Strand
233 bits(126)	1e-65	128/129(99%)	0/129(0%)	Plus/Minus

```
Query  15563  AAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGAA  15622
              ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  783    AAATTGCTGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGAA  724

Query  15623  ATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTCT  15682
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  723    ATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTCT  664

Query  15683  CAATGATGA  15691
              |||||||||
Sbjct  663    CAATGATGA  655
```

RaTG13 Spike Protein gene has 5% substitutions when comparing 2020 RNA-Seq and 2017 amplicon sequencing data.

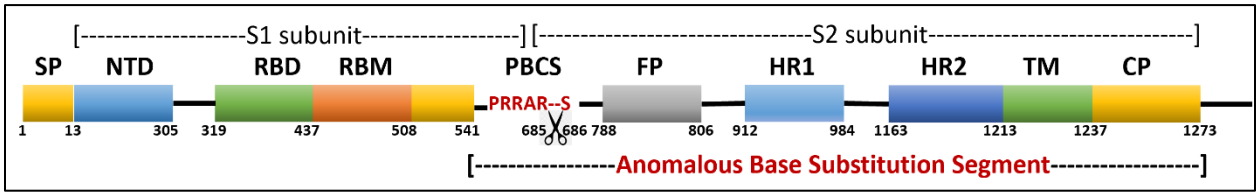
The segment of RaTG13 which shows the greatest sequence divergence between the RNA-seq and amplicon sequencing methods spans from A8886 to A9987 and is shown here below. It contains 80 base substitutions/indels in a 1107 nt sequence (5% substitution and 2% gaps).



No explanation has been offered in publications from the WIV for the method-dependent sequencing differences identified here, which are twenty- to 50-fold higher than the 0.1% technical error rate sometimes attributed to RNA-Seq data.

The Spike Protein gene sequence substitution divergence between RaTG13 and SARS-CoV-2 contains an improbable synonymous/non-synonymous pattern.

The functional structure of the SARS-CoV-2 Spike Protein is shown here:



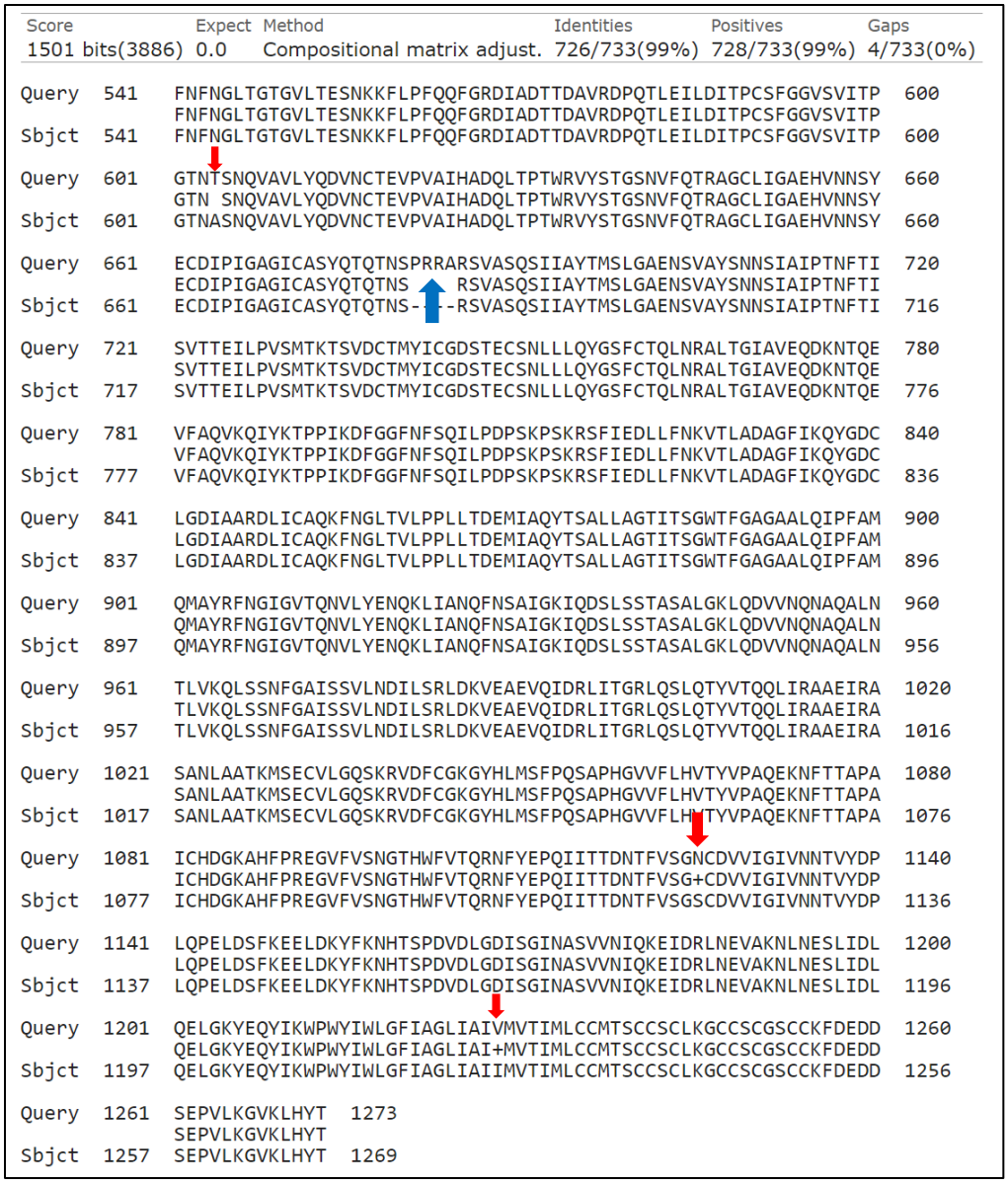
The SARS-CoV-2 Spike protein (above) contains an S1 subunit and S2 subunit with the Polybasic Cleavage Site (PBCS) between R685 and S686. This cleavage is performed by a host cell surface protease, furin, and is an important attribute in explaining the virulence of SARS-CoV-2 compared to other human coronaviruses, which do not have a furin cleavage site. The PBCS also contains the unusual PRRA insertion that has not been previously seen in Clade B coronaviruses and for which no natural mechanism for its appearance has been offered.[16]

The S1 subunit is located within the N-terminal 14–685 amino acids of S protein, containing N-terminal domain (NTD), receptor binding domain (RBD), and receptor binding motif (RBM). The S2 subunit contains a fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2), transmembrane domain (TM) and cytoplasmic domain (CP).

The base substitution pattern of synonymous and non-synonymous substitutions when comparing RaTG13 and the reference sequence of SARS-CoV-2 demonstrated an anomalous pattern for the coding region for aa 541 to 1273, a 733 aa protein segment representing over 60% of the SP gene.

As shown in the Text-Figure below, there are only three substitutions (red arrow) and the PBCS insertion (blue arrow) when comparing this segment of the RaTG13 and SARS-CoV-2 SP. Excluding the PBCS, the amino acid sequences are 99.6% identical.

[16] The proximal origin of SARS-CoV-2.



Score	Expect	Method	Identities	Positives	Gaps
1501 bits(3886)	0.0	Compositional matrix adjust.	726/733(99%)	728/733(99%)	4/733(0%)

```
Query  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600
             FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
Sbjct  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600

Query  601   GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
             GTN SNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
Sbjct  601   GTNASNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660

Query  661   ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
             ECDIPIGAGICASYQTQTNS   RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
Sbjct  661   ECDIPIGAGICASYQTQTNS-  -RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716

Query  721   SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
             SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
Sbjct  717   SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  776

Query  781   VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  840
             VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
Sbjct  777   VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  836

Query  841   LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  900
             LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM
Sbjct  837   LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  896

Query  901   QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  960
             QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN
Sbjct  897   QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  956

Query  961   TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1020
             TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Sbjct  957   TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1016

Query  1021  SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1080
             SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
Sbjct  1017  SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1076

Query  1081  ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP  1140
             ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSG+CDVVIGIVNNTVYDP
Sbjct  1077  ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDP  1136

Query  1141  LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL  1200
             LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
Sbjct  1137  LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL  1196

Query  1201  QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD  1260
             QELGKYEQYIKWPWYIWLGFIAGLIAI+MVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
Sbjct  1197  QELGKYEQYIKWPWYIWLGFIAGLIAIIMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD  1256

Query  1261  SEPVLKGVKLHYT  1273
             SEPVLKGVKLHYT
Sbjct  1257  SEPVLKGVKLHYT  1269
```

Given the high amino acid identity of this 733 amino acid sequence (except for the PBCS insertion) and the typical coronavirus synonymous to non-synonymous mutation frequency of between three and five synonymous mutations for each non-synonymous mutation,[17] it was expected that a comparison of the nucleotide sequence for this region between SARS-CoV-2 and RaTG13 would show an almost identical sequence as well.

In fact, when the SARS-CoV-2 nt sequence 23,183-25,384 was compared to the RaTG13 nt sequence 23,165-25,354, the corresponding genome sequence to the 99.6% identical protein sequence above, the nucleotide identity was only 94.2% identical, with 122 synonymous substitutions and only the three non-synonymous substitutions.

[17] Comparative genomic analysis

To put this in context a comparison of thirteen other protein coding regions of SARS-CoV-2 and RaTG13 (Text-Table below) shows that the overall synonymous to non-synonymous mutation frequency is 549 synonymous to 109 non-synonymous or a ratio of about 5.0.

Gene	Region of Genome	Total Nucleotides	Synonymous mutations	Non-Synonymous mutations	S/NS	Probability of more than the number of synonymous mutations given the probability of a synonymous mutation is 0.83 (based on all genes pooled)
pp1ab	1-21,239	21,239	659	102	6.5	0.003
pp1ab ABSS	**7448-18266**	**10,818**	**283**	**13**	**21.8**	**5.73 x 10^-12**
Spike Protein RBD	1-1814	1814	131	27	4.9	0.48
Anomalous Base Substitution Segment	**23,183-25,384**	**2201**	**112**	**3**	**37.3**	**< 1.0 x 10^-7**
Entire Spike Protein	1-3810	3808	231	41	5.6	0.18
ORF1a polyprotein	1-13,215	13215	440	86	5.2	0.33
ORF3a protein	1-828	828	25	6	4.2	0.56
E Protein	1-228	228	1	0	Infinite	0.83
M Protein	1-669	669	27	3	9.0	0.1
ORF6 Protein	1-186	186	3	0	Infinite	0.17
ORF7a Protein	1-366	366	13	3	4.3	0.47
ORF7b Protein	1-132	132	0	1	0	0.83
ORF8 Protein	1-366	366	5	6	0.8	0.99
Nucleocapsid Phosphoprotein	1-1260	1260	35	4	8.75	0.083

With the exception of the anomalous base substitution segment (ABSS) in the Spike Protein gene and the pp1ab gene, the remainder of the S/SN substitution ratios are consistent with the literature values for coronaviruses. Only two genes or gene regions have a higher S/SN ratio than the ABSS because they have no non-synonymous mutations: the E protein gene with 228 nucleotides and the ORF6 protein gene with 186 nucleotides. Because of the short length of these two genes, the probabilities of the results for the E and ORF6 genes were not significant, with p-values of 0.86 and 0.17, respectively.

The p-value for the ABSS, on the other hand, was highly significant, with a p-value of <0.0000001. This strongly suggests a non-natural cause for this base substitution pattern, barring some unknown biological mechanism for such a result.

A second highly anomalous sequence was found in the pp1ab gene. This is about five-times larger than the Spike Protein region and is even more unlikely to have happened naturally, a chance of about one in 100 billion times.

Are there only synonymous mutations in these regions because non-synonymous mutations lead to non-replicative viruses?

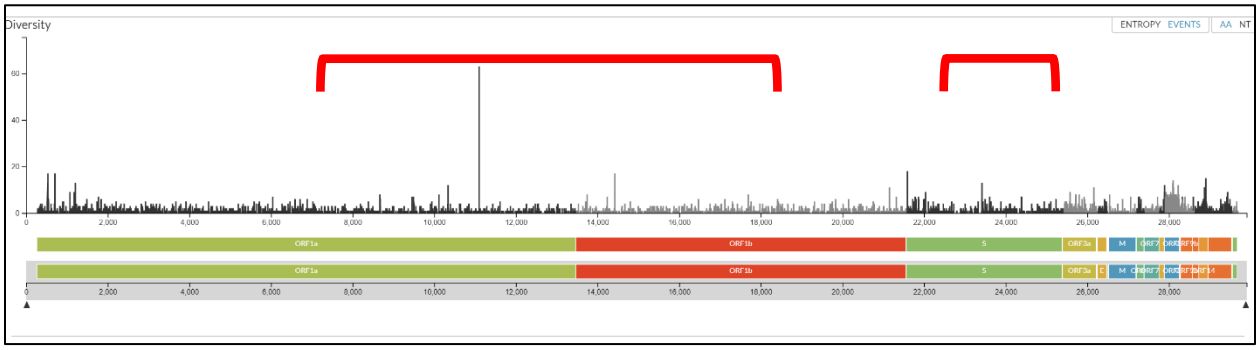
A simple explanation for these results would be an extreme criticality for the specific sequences of these regions with respect to infectivity. If a single amino acid change yielded a non-transmissible viral particle that strong negative purification process could explain the above results.

This hypothesis can be immediately rejected based on two observations.

In an examination of over 80,000 SARS-CoV-2 genome sequences, the most common Spike Protein non-synonymous mutation is within the ABSS (D614G) which was identified within weeks of the outbreak in January 2020 and which has become "the dominant virus…in every geographical region."[18] Specifically, as of August 28, 2020, GISAID reports that 65,738 full length SARS-CoV-2 genomes of a total of 83,387, or 79%, and comprising the G, GH, and GR clades, contain the D614G SNV. Under real world biological conditions, the ABSSN region has in fact, not a strong negative purification process in operation but in fact a strong positive selection process ongoing.

Secondly, in an analysis of mutations in 63,421 SARS-CoV-2 genomes the Spike Protein amino acid 605 to 1120 region had a total of 7,149 mutations. Fully 5,936 of these mutations (83%) are the above noted D614G non-synonymous change. Of the remaining 1213 mutations, 452 were non-synonymous while 755 were synonymous, a ratio of 1.7. There were also four indels and two stop codon mutations.

The following Text-Figure contains a map of the SARS-CoV-2 genome with the location of amino acid changes that have been found during the worldwide spread noted, with the frequency related to the height of the mark. The two ABSS in pp1ab and SP are marked with red brackets and clearly demonstrate an abundance of non-synonymous mutations in these regions during the human-to-human spread.



Nextstrain SARS-CoV-2 amino acid change events

Clearly, these regions can tolerate many non-synonymous mutations, rejecting the theory of a criticality for the amino acid sequence of this region. No other natural biological mechanism to explain these results has been identified.

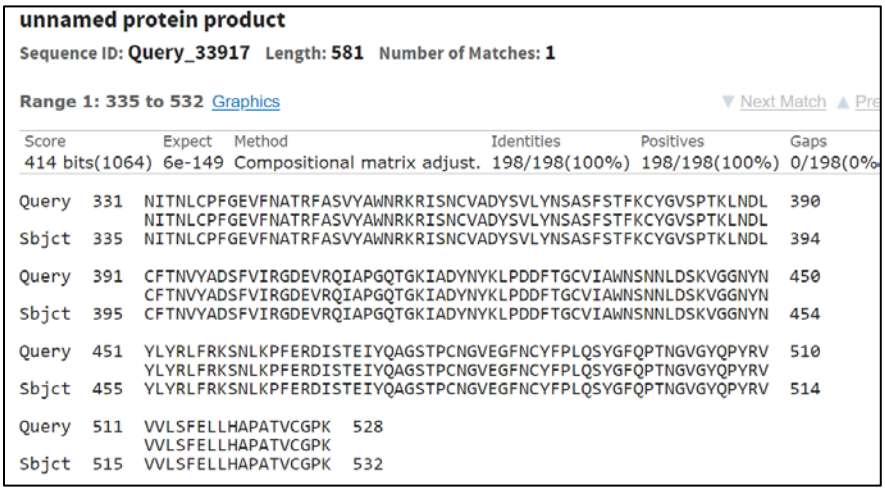
[18] Biswas NK, Majumder PP. Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. Indian J Med Res. 2020;151(5):450-458. doi:10.4103/ijmr.IJMR_1125_20.

Codon modification, enhancement, or optimization is an example from synthetic biology in which the S/SN ratio is, by design, an anomaly when looked at through the lens of nature

Synonymous codon substitution is a decades old, well known method of enhancing gene expression when cloning exogenous genes in a laboratory experiment. In a paper on the immunogenicity of the SARS-CoV-2 Spike Protein[19] the following synthetic biology methods were used:

"We used the following structure coordinates of the coronavirus spike proteins from the PDB to define the boundaries for the design of RBD expression constructs: SARS-CoV-2 (6VSB), SARS-CoV-1 (6CRV), HKU-1 (5I08), OC43 (6NZK), 229E (6U7H) NL63 (6SZS). Accordingly, a codon-optimized gene encoding for S1-RBD [SARS-CoV-1 (318 – 514 aa, P59594), SARS-CoV-2 (331 – 528 aa, QIS60558.1), OC43 (329 – 613 aa, P36334.1), HKU-1 (310 – 611 aa, Q0ZME7.1), 229E (295 – 433 aa, P15423.1) and NL63 (480 – 617 aa, Q6Q1S2.1)] containing human serum albumin secretion signal sequence, three purification tags (6xHistidine tag, Halo tag, and TwinStrep tag) and two TEV protease cleavage sites was cloned into the mammalian expression vector pαH. S1 RBDs were expressed in Expi293 cells (ThermoFisher) and purified from the culture supernatant by nickel-nitrilotriacetic acid agarose (Qiagen)."

The Genbank alignment (below) confirms that the authentic SARS-CoV-2 Spike Protein sequence (https://www.ncbi.nlm.nih.gov/nuccore/1798174254) and the Synthetic construct SARS CoV-2 spike protein receptor binding domain gene, complete cds are 100% homologous at the protein level:



unnamed protein product

Sequence ID: **Query_33917** Length: **581** Number of Matches: **1**

Range 1: 335 to 532 Graphics ▼ Next Match ▲ Pre

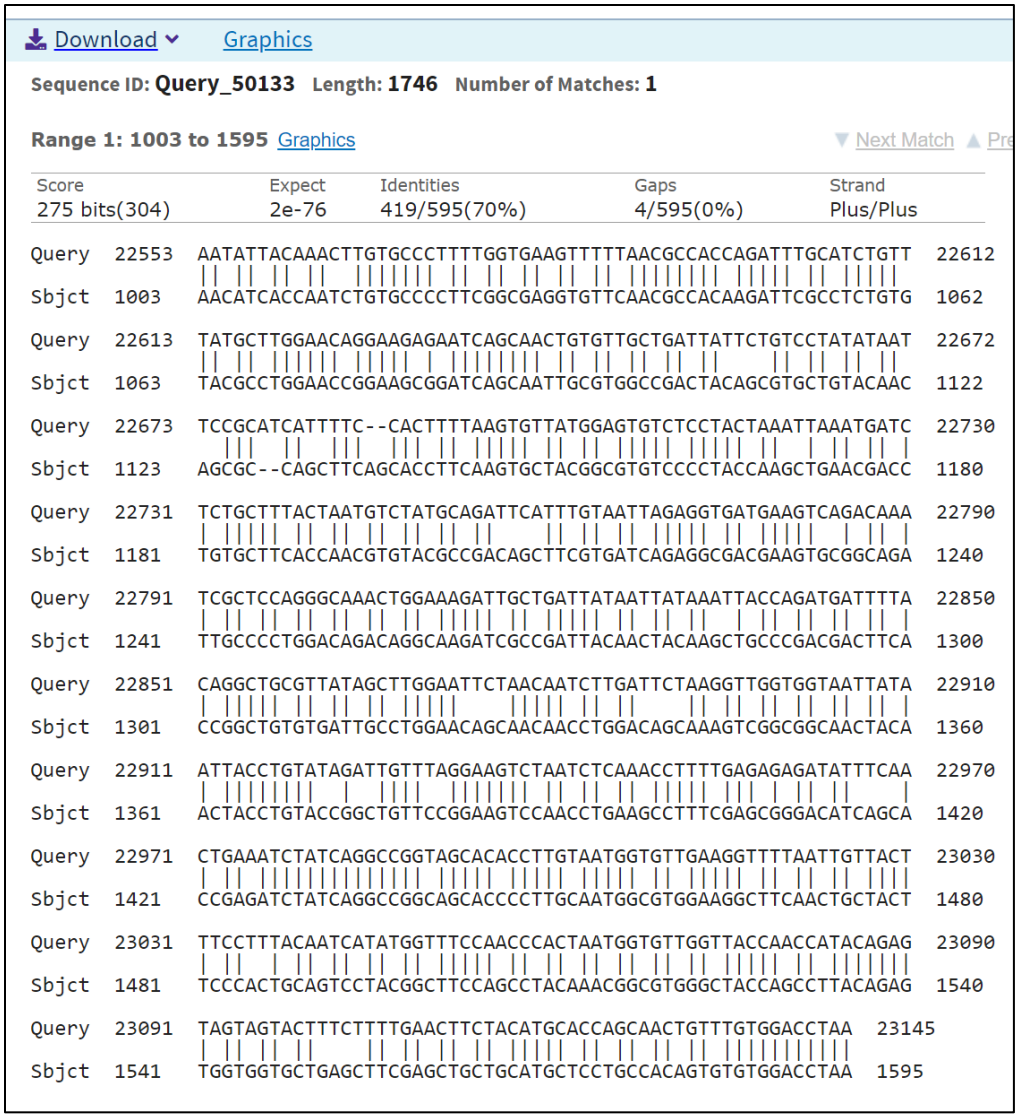
Score	Expect	Method	Identities	Positives	Gaps
414 bits(1064)	6e-149	Compositional matrix adjust.	198/198(100%)	198/198(100%)	0/198(0%

```
Query  331  NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL  390
            NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL
Sbjct  335  NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL  394

Query  391  CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN  450
            CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
Sbjct  395  CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN  454

Query  451  YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV  510
            YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
Sbjct  455  YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV  514

Query  511  VVLSFELLHAPATVCGPK  528
            VVLSFELLHAPATVCGPK
Sbjct  515  VVLSFELLHAPATVCGPK  532
```

But a comparison of the authentic nucleotide sequence of SARS-CoV-2 to the codon-optimized synthetic construct shows no match using the "highly similar Megablast" algorithm setting. When the alignment algorithm is run in a more relaxed mode the impact of codon optimization in this case can be seen, a 70% homology:

[19] https://immunology.sciencemag.org/content/5/48/eabc8413/tab-pdf

The RaTG13 fecal specimen appears contrived,
genome assembly inaccurate, and lab synthetic biology signature apparent



Download Graphics

Sequence ID: Query_50133 **Length:** 1746 **Number of Matches:** 1

Range 1: 1003 to 1595 Graphics Next Match Pre

Score	Expect	Identities	Gaps	Strand
275 bits(304)	2e-76	419/595(70%)	4/595(0%)	Plus/Plus

```
Query  22553  AATATTACAAACTTGTGCCCTTTTGGTGAAGTTTTTAACGCCACCAGATTTGCATCTGTT  22612
              || || || ||  ||||||| || || || || || |||||||| ||||| || |||||
Sbjct  1003   AACATCACCAATCTGTGCCCCTTCGGCGAGGTGTTCAACGCCACAAGATTCGCCTCTGTG  1062

Query  22613  TATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAAT  22672
              || || |||||| ||||| | |||||||| || || || || ||    || || || ||
Sbjct  1063   TACGCCTGGAACCGGAAGCGGATCAGCAATTGCGTGGCCGACTACAGCGTGCTGTACAAC  1122

Query  22673  TCCGCATCATTTTC--CACTTTTAAGTGTTATGGAGTGTCTCCTACTAAATTAAATGATC  22730
                 |||  ||  |||  ||| || ||||| || || ||||| ||||| ||  | || || |
Sbjct  1123   AGCGC--CAGCTTCAGCACCTTCAAGTGCTACGGCGTGTCCCCTACCAAGCTGAACGACC  1180

Query  22731  TCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAAGTCAGACAAA  22790
              | ||||| || || || || || ||    || || || ||||| || |||||  | || |
Sbjct  1181   TGTGCTTCACCAACGTGTACGCCGACAGCTTCGTGATCAGAGGCGACGAAGTGCGGCAGA  1240

Query  22791  TCGCTCCAGGGCAAACTGGAAAGATTGCTGATTATAATTATAAATTACCAGATGATTTTA  22850
              | || || || || || || ||||| || ||||| || || ||  | || || || || |
Sbjct  1241   TTGCCCCTGGACAGACAGGCAAGATCGCCGATTACAACTACAAGCTGCCCGACGACTTCA  1300

Query  22851  CAGGCTGCGTTATAGCTTGGAATTCTAACAATCTTGATTCTAAGGTTGGTGGTAATTATA  22910
              | ||||| || || || |||||    ||||| || ||    || || || || || || |
Sbjct  1301   CCGGCTGTGTGATTGCCTGGAACAGCAACAACCTGGACAGCAAAGTCGGCGGCAACTACA  1360

Query  22911  ATTACCTGTATAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAA  22970
              | ||||||||  |  ||||  ||||||| || || || ||||| ||| | || ||    |
Sbjct  1361   ACTACCTGTACCGGCTGTTCCGGAAGTCCAACCTGAAGCCTTTCGAGCGGGACATCAGCA  1420

Query  22971  CTGAAATCTATCAGGCCGGTAGCACACCTTGTAATGGTGTTGAAGGTTTTAATTGTTACT  23030
              | || |||||||||||||| ||||| ||||| ||||| || ||||| || || || ||||
Sbjct  1421   CCGAGATCTATCAGGCCGGCAGCACCCCTTGCAATGGCGTGGAAGGCTTCAACTGCTACT  1480

Query  23031  TTCCTTTACAATCATATGGTTTCCAACCCACTAATGGTGTTGGTTACCAACCATACAGAG  23090
              | ||  | || || || || ||||| || || || || || || ||||| || |||||||
Sbjct  1481   TCCCACTGCAGTCCTACGGCTTCCAGCCTACAAACGGCGTGGGCTACCAGCCTTACAGAG  1540

Query  23091  TAGTAGTACTTTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAA  23145
              | || || ||    || || || || ||||| || || || || ||||||||||
Sbjct  1541   TGGTGGTGCTGAGCTTCGAGCTGCTGCATGCTCCTGCCACAGTGTGTGGACCTAA  1595
```

This is a situation in which there are 176 synonymous changes without a single non-synonymous change and is the genome signature of laboratory-derived synthetic biology. If these sequences were compared for phylogenetic divergence without the knowledge of their artificial construction, this synthetic laboratory experiment would create the impression that these two sequences had diverged in the wild from a common ancestor decades earlier.

The following Table identifies four regions of the RaTG13 and SARS-CoV-2 genomes in which there were a total of 220 synonymous mutations without a single non-synonymous change.

Protein/Gene	Protein Region	Total Nucleotides	Synonymous mutations	NS Mutations
S Protein	605-1124	1557	91	0
pp1ab	3607-4534	2781	66	0
pp1ab	4626-5111	1455	26	0
pp1ab	5113-5828	2145	37	0
	Total	7938	220	0

These regions represent over 26% of the entire genome and appear analogous to the outcome expected from the application of a synonymous codon modified, laboratory-derived synthetic biology project. They also represent about one-sixth of the 4% apparent phylogenetic divergence between RaTG13 and SARS-CoV-2.

Discussion. The foundation of the working hypothesis that the COVID-19 pandemic arose via a natural zoonotic transfer from a non-human vertebrate host to man has been built on two publications: the February 3, 2020 *Nature* paper by Dr. Zheng-Li Shi and colleagues, in which the bat coronavirus RaTG13 is first identified as the closest sequence identity to SARS-CoV-2 at 96.2% and the March 17, 2020 *Nature Medicine* paper entitled, "The proximal origin of SARS-CoV-2," by Andersen *et al.,* in which the Shi *et al.* paper is cited as evidence for a bat origin for the pandemic. In the approximately six months since they were published, these two papers have been cited over 1600- and 200-times on PubMed, respectively.

However, research is beginning to question whether a bat species can be considered a natural reservoir for SARS-CoV-2. A recent paper performed an *in silico* simulation of the SARS-CoV-2 Spike Protein interaction with the cell surface receptor, ACE2, from 410 unique vertebrate species, including 252 mammals.[20] Among primates, 18/19 have an ACE2 receptor which is 100% homologous to the human protein in the 25 residues identified to be critical to infection, including the *Chlorocebus sabaeus* (the Old World African Green monkey) and the rhesus macaques.

It is noteworthy that the laboratory workhorse of coronavirus research is the VERO cell, isolated from a female African Green monkey in 1962, and containing an ACE2 receptor that is 100% homologous to the human ACE2 in the 25 critical amino acids for infectivity.

This *in silico* work was confirmed in the laboratory with respect to rhesus macaques. Within weeks of the identification of SARS-CoV-2, the Wuhan laboratory had demonstrated that the pandemic virus would infect and produce a pneumonia in rhesus macaques.[21]

A surprising finding from the ACE2 *in silico* surveillance work was the very poor predicted affinity of the ACE2 receptors in both bats and pangolins. Of 37 bat species studied, 8 scored low and 29 scored very low. As expected by these predictions, cell lines derived from big brown bat

[20] Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates Joana Damas, et al. Proc. of the Nat. Acad. of Sci. Aug 2020, 202010146; DOI: 10.1073/pnas.2010146117

[21] Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the *Rhesus Macaques.* C. Shan et al., Research Square, **DOI:** 10.21203/rs.2.25200/v1. Shan, C., Yao, Y., Yang, X. *et al.* Infection with novel coronavirus (SARS-CoV-2) causes pneumonia in *Rhesus macaques. Cell Res* **30,** 670–677 (2020). https://doi.org/10.1038/s41422-020-0364-z

(*Eptesicus fuscus*),[22] Lander's horseshoe bat (*Rhinolophus landeri*), and Daubenton's bat (*Myotis daubentonii*) could not be infected with SARS-CoV-2.[23]

It is unfortunate that growth of the RaTG13 specimen could not have been attempted in the *Rhinolophus sinicus* primary or immortalized cells generated and maintained in the Wuhan laboratory: kidney primary cells (RsKi9409), lung primary cells (RsLu4323), lung immortalized cells (RsLuT), brain immortalized cells (RsBrT) and heart immortalized cells (RsHeT).[24] However it should be noted that a synthetically created RaTG13 was reported not to infect human cells expressing *Rhinolophus sinicus* ACE2, providing evidence that RaTG13 may not be a viable coronavirus in a wild bat population.[25]

The other proposed intermediate host, the pangolin, also had predicted ACE-2 affinity that was either low or very low.

A recent paper that examined the high synonymous mutation difference between RaTG13 and SARS-CoV-2 used an *in silico* methodology to suggest that the difference could be largely attributed to the RNA modification system of hosts.[26] However, the authors do not "(t)he limitation of our study is that we were currently unable to provide experimental evidence for the modification on viral RNAs." The low S/SN ratio of 1.7 in the expansion of SARS-CoV-2 in the human population would argue against a robust host RNA modification mechanism.

In summary, the findings reported here are:

1. Inconsistences between published papers and interviews as to the source and sequencing history of the original specimen that was claimed to have been collected in 2013 (RaBtCoV/4991) and the specimen for the bat RaTG13 virus. For example, two explanations of the discovery of the close relationship between RaTG13 and SARS-Cov-2, a highly homologous match between the RdRp genes of the viruses noticed in 2020 followed by full genome sequencing, or identification in 2020 of a homologous match to full genome sequencing previously done in 2018. Current publicly available data for RaTG13 from 2017 and 2018 is a set of 33 amplicon sequencing runs but they cover only about 80% of the entire genome. In the *Science* interview Dr. Shi's says the specimen for RaTG was consumed during sequencing in 2018, but if this is true, the RNA-Seq referred to in the *Nature* paper could not have been performed in 2020. At this time, the Wuhan

[22] J. Harcourt et al., Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. Emerg. Infect. Dis. 26, 1266–1273 (2020).
[23] M. Hoffmann et al., SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 271–280.e8 (2020).
[24] Zhou, P., Fan, H., Lan, T. et al. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature 556, 255–258 (2018). https://doi.org/10.1038/s41586-018-0010-9.
[25] Y. Li et al., Potential host range of multiple SARS-like coronaviruses and an improved ACE2-Fc variant that is potent against both SARS-CoV-2 and SARS-CoV-1. bioRxiv:10.1101/2020.04.10.032342 (18 May 2020).
[26] The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification

laboratory has not met the requirements of *Nature* with respect to the sharing of primary and sequence assembly data from their seminal paper[1] and this data should be provided immediately.

2. The specimen from which RaTG13 was reported to have been isolated and which has been repeatedly reported to have been a bat fecal specimen has a taxonomical composition of eukaryotes, bacteria, and viruses that is completely different from a set of nine bat fecal specimens collected in the same field visits by the same laboratory personnel from the Wuhan Institute of Virology. The probability that an authentic fecal specimen could have the composition reported is one in ten million, an impossibly low occurrence. Examination of the strong signals in the RaTG13 specimen identifies both a variety of bat genetic material, some that are not native to China, as well as unexpected species, such as marmots and a red fox. It also contains a telltale 3% primate sequence consistent with VERO cell contamination. I propose that this specimen is apparently either a mislabeled specimen (although I cannot conjure what the field source or specimen would be) or was artificially created in a laboratory.
3. The method-dependent sequence differences between the amplicon data and the RNA-Seq data are about 5% or about 50-times higher than expected as a technical error rate of 0.1%. This is an experimental quality issue that needs to be addressed; no explanation has been offered for this to date. In addition, no assembly methodology has been provided and at least two gaps, totaling over 60 nt, were easily identified.
4. The findings, reported here of a mutational drift of synonymous mutations only between SARS-CoV-2 and RaTG13 in the Spike Protein S1/S2 region and the pp1ab gene that has never been seen in nature before and which has a probability of having occurred by chance of less than one in ten million and one in one billion makes it more likely that, at least for these portions of the RaTG13 genome, comprising over one-quarter of the entire genome, another process is underway. With the demonstration that codon-enhancement or optimization can produce this unnatural S/SN pattern, some form of laboratory-based synthetic biology was performed on RaTG13, SARS-CoV-2, or both.

Apparently, the entire specimen from which RaTG13 was purported to have been found has been consumed in previous sequencing experiments and the Principal Investigator has stated that no virus has ever been isolated or cultured from the specimen at any time in the past. Given the irregularities and anomalies identified in this paper it seems prudent to conclude that all data with respect to RaTG13 must be considered suspect. As such, reliance of the foundational papers of the origin of SARS-CoV-2 as having arisen from bats via a zoonotic mechanism must be reexamined and questioned.

FINANCIAL DISCLOSURE. The author is CEO of Atossa Therapeutics, Inc. (NASDAQ: ATOS), a clinical-stage biopharmaceutical company seeking to discover and develop innovative medicines in areas of significant unmet medical need. Atossa's current focus is on breast cancer and COVID-19 therapeutics. The author received no funding for this research from any source.