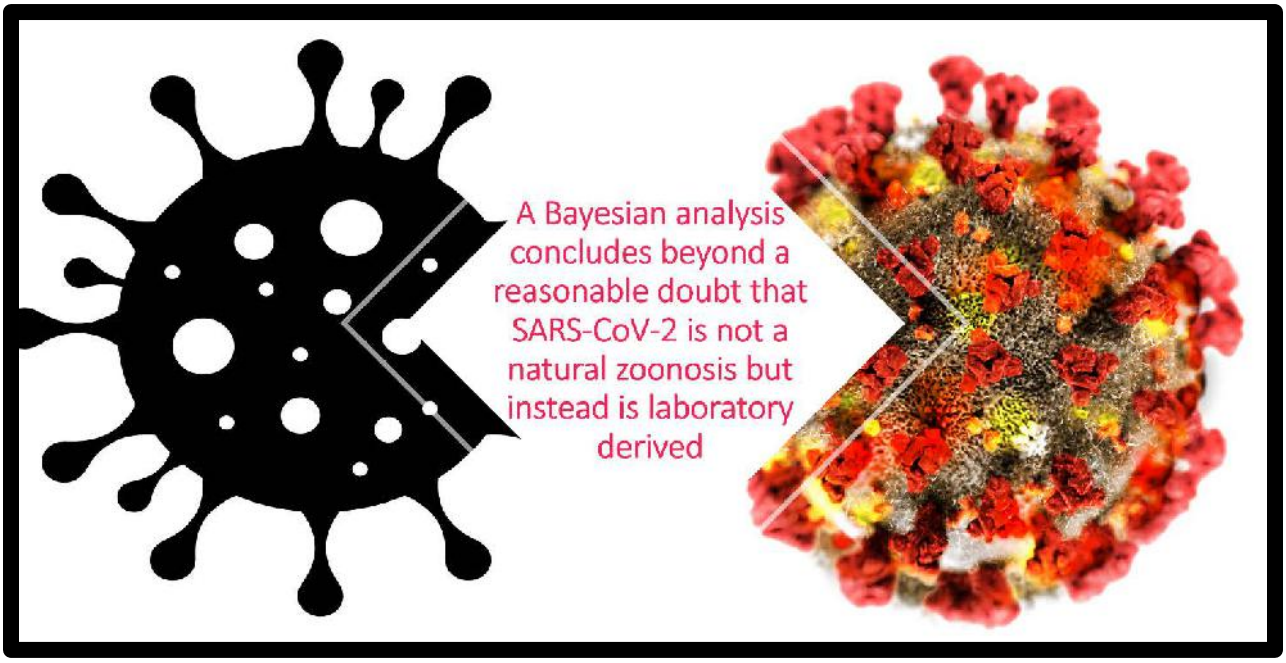


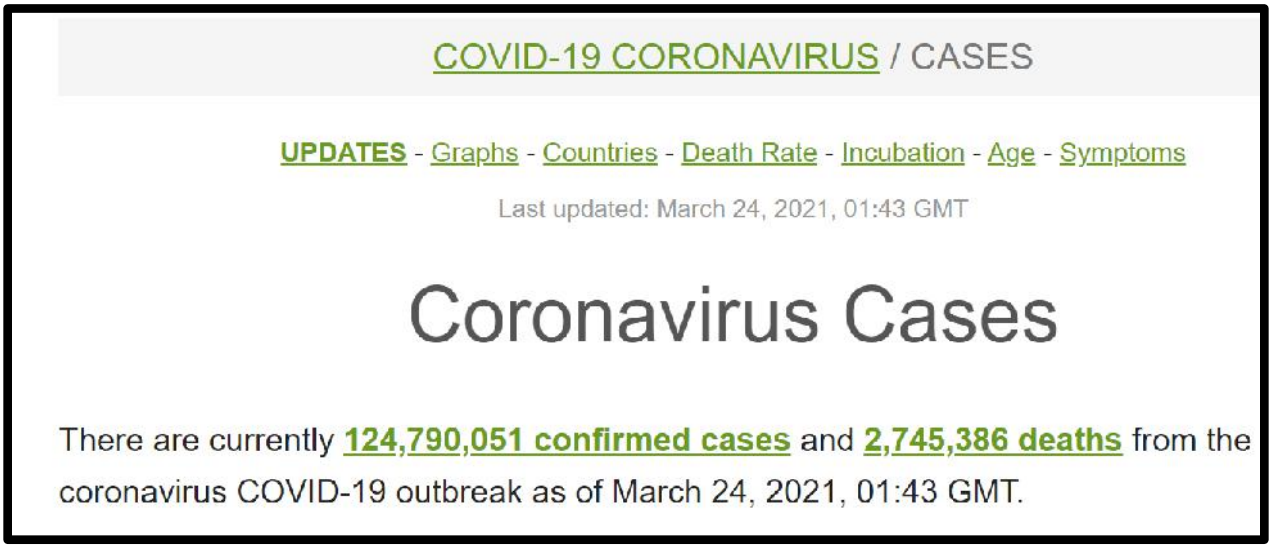
By

Steven Carl Quay, MD, PhD

www.DrQuay.com

Steven@DrQuay.com

Dedication. This work is dedicated to the men, women, and children who were infected with SARS-CoV-2 over the last year. It is my hope that this work becomes part of the body of evidence to help inform the public about gain-of-function pathogen research and that a renewed debate can be had about the benefits and risks of this research in the context of world health.



Source: https://www.worldometers.info/coronavirus/

Acknowledgements. Despite having collaborated over many decades on numerous scientific projects, research during 2020 into COVID-19, SARS-CoV-2, and therapeutic approaches has been a unique experience. With lockdowns and international travel bans, all collaborative work has been virtual. With an apparent bias surrounding investigation into the origin of SARS-CoV-2, ad hoc groups of Citizen-Scientists, often anonymous, have worked together via email, videoconference, micro-blogging, and social messaging networks to advance our understanding of this horrific pandemic.

I want to thank a Twitter group called #DRASTIC for many useful discussions that found their way into this document. Dr. Martin Lee, Ph.D., Adjunct Professor of Statistics at UCLA provided statistical support throughout this work. I want to thank D.A. for originally suggesting performing a Bayesian analysis on the work I had done on SARS-CoV-2 and for his facilitation of the review of this work by a diverse group of scientists and policy makers.

In all cases, however, this is my own work product.

A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

Executive Summary. The one-year anniversary of the COVID-19 pandemic records 2.1 million deaths, over 100 million confirmed cases,[1] and trillions of dollars of economic damage. Although there is universal agreement that a coronavirus identified as Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 (abbreviated CoV-2 henceforth) causes the disease COVID-19, there is no understanding or consensus on the origin of the disease.

The Chinese government, WHO, media, and many academic virologists have stated with strong conviction that the coronavirus came from nature, either directly from bats or indirectly from bats through another species. Transmission of a virus from animals to humans is called a zoonosis.

A small but growing number of scientists have considered another hypothesis: that an ancestral bat coronavirus was collected in the wild, genetically manipulated in a laboratory to make it more infectious, training it to infect human cells, and ultimately released, probably by accident, in Wuhan, China. For most of 2020 this hypothesis was considered a crackpot idea, but in the last few weeks, more media attention has been given to the possibility that the Wuhan Institute of Virology, located near the Wuhan city center and with a population of over 11 million inhabitants, may have been the source of the field specimen collection effort, laboratory genetic manipulation, and subsequent leak. On January 15, 2021, the U.S. Department of State issued a statement requesting the WHO investigation of the origin of COVID-19 include specific assertions related to a laboratory origin of the pandemic.[2]

Given the strong sentiment in the scientific community in favor of a zoonosis and the massive effort undertaken by China to find the natural animal source, one can assume that any evidence in favor of a natural origin, no matter how trivial, would become widely disseminated and known. This provides a potential evidence bias within the scientific community in favor of a natural origin which isn't quantifiable but should be kept in mind.

This becomes especially important background when evidence that could support a laboratory origin has been directly provided by leading Chinese scientists themselves, like Dr. Zhengli Shi, head of coronavirus research at the Wuhan Institute of Virology and Gao Fu (George Fu Gao), Director of Chinese CDC; by the Chinese government, as well as by powerful and vocal, pro-natural origin scientists, like Dr. Peter Daszak, of the NYC-based NGO, EcoHealth Alliance.

This report uses Bayesian inference, a common statistical tool in which Bayes' theorem, a well-known statistical equation, is used to update the likelihood for a particular hypothesis as more evidence or information becomes available. It is widely used in the sciences and medicine and has begun to be used in the law.

[1] https://www.worldometers.info/coronavirus/?
[2] https://www.state.gov/ensuring-a-transparent-thorough-investigation-of-covid-19s-origin/

The starting probability for origin of SARS-CoV-2 was set with the zoonotic or natural hypothesis at 98.8% likelihood with the laboratory origin hypothesis set at 1.2%. The initial state was biased as much as possible towards a zoonotic origin, with the starting point selected as the upper bounds of the 95% confidence interval for the mean and standard deviation of three independent estimates, including one by Daszak and colleagues. Each piece of new evidence for or against each hypothesis was then used to adjust the probabilities. If evidence favored a natural origin the math adjusts upward the probability of a natural origin, and so on.

The outcome of this report is the conclusion that the probability of a laboratory origin for CoV-2 is 99.8% with a corresponding probability of a zoonotic origin of 0.2%. This exceeds most academic law school discussions of how to quantify 'beyond a reasonable doubt,' the threshold for finding guilt in a criminal case. The report contains the detailed analysis and quantitative basis for the statistics and conclusion. It should be noted that because of the commutative property of the collected adjustments to the probabilities, the order in which they are used in the overall calculation is immaterial and the same end likelihoods will be reached regardless of the order of input.

The following Text-Table summarizes the evidence examined and the changes in probabilities:

Evidence	Zoonotic Origin	Laboratory Origin
Initial State	98.8%	1.2%
International committees to determine CoV-2 origin may not be impartial	98.8%	1.2%
Three key zoonotic papers: pros and cons	98.8%	1.2%
SARS-like infections among employees of the Wuhan Institute of Virology in the fall of 2019 reported by US Government	98.8%	1.2%
Location of first cases near Wuhan Institute of Virology	95.1%	4.9%
Lack of evidence of seroconversion in Wuhan and Shanghai	80.9%	19.1%
Lack of posterior diversity	30.8%	69.2%
Opportunity: The Wuhan Institute of Virology has publicly disclosed that by 2017 it had developed the techniques to collect novel coronaviruses, systematically modify the receptor binding domain to improve binding or alter zoonotic tropism and transmission, insert a furin site to permit human cell infection, make chimera and synthetic viruses, perform experiments in humanized mice, and optimize the ORF8 gene to increase human cell death.	30.8%	69.2%
Lack of furin cleavage sites in any other sarbecovirus	4.7%	95.3%
Rare usage of -CGG- single codons & no CGG-CGG pairs	0.5%	99.5%
Routine use of CGG in laboratory codon optimization, including Daszak & Shi	0.2%	99.8%
Spike Protein receptor binding region (200 amino acids) optimized for humans	0.2%	99.8%
Whole genome analysis shows pre-adaption of CoV-2	0.2%	99.8%
The finding of CoV-2 in Barcelona wastewater in early 2019 was an artifact	0.2%	99.8%
Shi and the WHO comment early on that CoV-2 seemed to begin with a single patient	0.2%	99.8%
Mammalian biodiversity between Yunnan and Hubei is significantly different, limiting a potential common intermediate host	0.2%	99.8%
The ancestor of CoV-2 can only obtain a furin site from other subgenera viruses but recombination is limited/non-existent between subgenera	0.2%	99.8%
Canvas of 410 animals shows humans and primates are the best, bats are the worst, for ACE2-Spike Protein interaction	0.2%	99.8%
A government requested review of samples collected from a mineshaft may have caused the COVID-19 pandemic	0.2%	99.8%
The Hunan Seafood Market and farmed animals in Hubei province are not the source of CoV-2	0.2%	99.8%
Line 2 of the Wuhan Metro System is the likely conduit of the pandemic and is the closest subway line to the WIV	0.2%	99.8%
Feral and domestic cats are not the intermediate host	0.2%	99.8%
Extraodinary pre-adaption for the use of human tRNA is observed	0.2%	99.8%
Evidence of lax operations and disregard of laboratory safety protocols and regulations in China	0.2%	99.8%
Previous SARS-CoV-1 laboratory accidents	0.2%	99.8%
Shi and Daszak use Wuhan residents as negative control for zoonotic coronavirus exposure	0.2%	99.8%
RaTG13 could be CoV-2 precursor using the synthetic biology 'No See 'Em' technique	0.2%	99.8%
Location, location, location: Based on the distance between known SARS-CoV-1 laboratory-acquired infections and the hospital of admission of the infected personnel, the WIV is within the expected hospital catchment for a CoV-2 LAI	0.2%	99.8%

The summary which follows will simply be a review and discussion of the evidence in the context of the two hypotheses.

Zoonosis Hypothesis

A viral zoonosis has at least three elements, a host, a virus, and the human population. With some viruses there are often two hosts. One is a 'reservoir host' where the virus can live for years or even decades in a relatively stable relationship. The reservoir host is never decimated by the virus, and the virus is never burned out by the reservoir host, disappearing completely. For coronaviruses the reservoir host is always one or more bat species. If there is a reservoir host that some viruses that cannot jump directly into the human population, there is a need for an second host, an intermediate host. In this case the virus spends time jumping into the intermediate host, 'practicing' adaption through random mutation and Darwinian selection for fitness to reproduce, infect, and transmit in the intermediate host. This process is then repeated between the intermediate host and the human population. Alternatively, the virus can jump directly between the bat reservoir and humans, without the need for an intermediate host.

For two prior human coronavirus epidemics, an intermediate or proximate host was identified. For SARS-CoV-1 in 2003-4 it was the civet cat while for Middle Eastern Respiratory Syndrome (MERS) in 2012-4 it was the camel. In both of these human epidemics, the intermediate host was identified within four to ten months of the first clinically identified human infection. With CoV-2 we are at 12 months since the pandemic began and still waiting for evidence of, despite a much larger effort inside China to find an intermediate host. For both of these previous pandemics, a bat species reservoir host was also identified, but not in the case of SARS-CoV-2.[3]

Based on the genome sequence of CoV-2, Drs. Shi and Daszak have proposed that the reservoir host for CoV-2 is the intermediate horseshoe bat (*Rhinolophus affinis*), which is found in Yunnan Province. Yunnan Province is in southern, rural China and about 1900 km from the north central province of Hubei, where the 11 million people of Wuhan live. In the US this would be equivalent in distance, climate change, and human population density difference to going from the Everglades in Florida to Manhattan, in New York City. The intermediate horseshow bat isn't found at all in Hubei province, making a direct bat-to-human transmission improbable.[4] Experiments in three independent laboratories also demonstrate that CoV-2 has changed genetically so much that it can no longer infect any bat species cell culture tested. So, while the leading US coronavirus expert, Dr. Ralph Baric of The University of North Carolina suggested in early 2020 that CoV-2 may have jumped into the human population directly from bats without an intermediate host, this hypothesis seems to no longer be viable.

For the zoonosis hypothesis to be advanced, it is now necessary to find an intermediate host. In January 2020 a theory was proposed that CoV-2 arose in the Huanan Seafood Market, a

[3] I am distinguishing here the difference between SARS-CoV-2 being a descendent of a bat coronavirus (with 3.8% or 1100 nucleotide (nt) differences between them) and the finding of the immediate precursor of SARS-CoV-2 in a bat colony population somewhere in the wild, which usually is <100 nt differences.

[4] "We have done bat virus surveillance in Hubei Province for many years but have not found that bats in Wuhan or even the wider Hubei Province carry any coronaviruses that are closely related to SARS-CoV-2. I don't think the spillover from bats to humans occurred in Wuhan or in Hubei Province," said Dr. Shi. Science, July 2020

traditional Chinese "wet market" where live animals are butchered and sold for food. The market theory was based on the observation that about 40% of early patients worked or shopped there. This was reminiscent of the wet market sources for civet cats infected with SARS-CoV-1 or the camel markets for the MERS coronavirus. The Chinese authorities closed the market on December 31, 2019 after performing extensive environmental sampling and sanitation.

But by May 2020 Dr. Gao Fu, Director of the Chinese CDC, announced that the market was not the source of CoV-2, as all of the animal specimens tested negative for CoV-2. And while SARS-CoV-1 was found in 100% of local farmed civets when tested, CoV-2 was different. In July 2020 Dr. Shi reported that extensive testing of farmed animals throughout Hubei Province failed to find CoV-2 in any animals.

For about six months, the pangolin, a scaly solitary-living anteater, was suspected to be the intermediate host but finally Dr. Daszak reported that CoV-2 was not found in pangolins in the wild or from the (illegal) market trade.[5] Domestic and feral cats also were ruled out as a possible source. A comprehensive computer-based screen of 410 different animals reported the remarkable finding that the best ACE2 receptor matches to CoV-2 were human and other primates (or primate cells in the laboratory), including the favorite laboratory coronavirus host, the VERO monkey cell culture, and that all bat species were the worst host. At the time of this writing, there is not even a working hypothesis for the species of an intermediate host.

A typical zoonosis has a number of characteristic properties that can allow identification of a zoonotic infection, even in the absence of identifying an intermediate host. None of these properties are found for CoV-2.

All zoonotic infections have in common the principle that when a virus in nature uses evolution to move from, for example, a bat host to a camel host and then to a human host, it is a hit and miss, slow process. After all, evolution is the result of random genetic changes, mutations, and then enrichment of the ones that are helpful by amplification during reproduction. With both SARS-CoV-1 and MERS, the coronavirus spent months and years jumping from the intermediate host into humans, not having all of the necessary mutations needed to be aggressive, grow, and then spread, but spending enough time in humans to cause an infection and leaving behind a corresponding immune response.

The hallmark evidence of this 'practice' in abortive host jumping is in stored, archived human blood specimens taken from before the epidemic, where one can find evidence of pre-epidemic, usually sub-clinical, community spread from the antibodies to the eventual epidemic virus. For SARS-CoV-1 and MERS, about 0.6% of people in the region where the epidemic began showed signs of an infection in archived blood. With CoV-2, this seroconversion, as it is called, has never been observed, including in 540 specimens collected from 'fever clinics' in Wuhan between October 2019 and January 2020, reported by the WHO. Because this is such a potent signal of a zoonosis, and because I believe that China has over 100,000 stored specimens from

[5] https://link.springer.com/article/10.1007/s10393-020-01503-x

Wuhan taken in the fall of 2019, the lack of reports of seroconversion, the silence from China on this evidence, speaks volumes.

Another hallmark of a slow, natural zoonosis can be found in the virus. In SARS-CoV-1 and MERS, the coronavirus spent years in the intermediate host, passing back and forth among populations of hosts, the civets or camels, that were living in close proximity. During this time, they would accumulate a background of genetic mistakes, i.e., mutations- usually about one mistake every two weeks. When the final chip falls, and a mutation(s) happens allowing the jump into humans, the virus with that new mutation(s) also jumps around within the intermediate host population. The consequence of this latter behavior for a true zoonosis is that the genome sequences found in humans don't all descend from a single jump into a single human but show jumps from viruses that are only cousins of each other, not direct lineal descendants.

In a true zoonosis, the family tree of virus genome sequences doesn't pass back through the first patient but instead tracks all the way back to an ancestor months or years earlier. This is called posterior diversity, and it is an easy genetic test to perform. With CoV-2, every one of the more than 294,000 virus genomes sequenced can be traced back to the first genomic cluster and in the first patient in that cluster, a 39-year-old man who was seen at the People's Liberation Army (PLA) Hospital about one mile from the Wuhan Institute of Virology. The CoV-2 pandemic has the phylogenetic signature of one pure virus sequence infecting one human, with human-to-human spread thereafter; there is just the one and only jump into the human population ever seen. This lack of posterior diversity has been alluded to by Dr. Shi, by the WHO, and by other prominent virologists; they just never take that critical piece of the evidence to the next the proper inference.

The virus in a true zoonosis also contains the signature record of the gradual changes and adaptions it made in the protein key, the Spike Protein, it uses to unlock human cells and cause infection. With SARS-CoV-1 the Spike Protein had fewer than one-third of all the changes it would later develop by the time it became an epidemic. With CoV-2 the Spike Protein was almost perfectly adapted to the human lock, using 99.5% of the best amino acids possible.

Since with CoV-2 we have no evidence from stored blood that it was quietly practicing on humans in the community of Wuhan, it is surprising that when it finds its first patient, it has perfected to 99.5% the spike protein amino acid sequence, its ability to attack and infect humans. If this adaption couldn't have happened in the community, the only place it could have happened is in a laboratory, by what is called serial passage, a common laboratory process that repeatedly gives the virus a chance to practice on humanized mice or VERO monkey cells.[6] A related study showing human adaption right from the start of the pandemic looked at which of the dozens of protein manufacturing tools that CoV-2 uses (called tRNAs). It showed the same uncanny adaptation to the human tools with no evidence that the tools from other potential intermediate hosts would be suitable.

[6] It is noteworthy that the furin cleavage site is actually unstable in passage in VERO cells and is often deleted within a few passages. A laboratory origin theory needs to account for this observation. On the other hand, mutations in the furin site among the human CoV-2 genomes are exceedingly rare.

This evidence presented makes a strong case that CoV-2 did not come from nature. But is there affirmative evidence that it could have come from a laboratory? The answer is yes.

Laboratory Origin Hypothesis

The spike protein that gives the coronavirus its name, corona or crown, is the key to match with the lock found in host cells. But before it can inject its genetic material in the host cell, the spike protein needs to be cut, to loosen it in preparation for infection. The host cell has the scissors or enzymes that do the cutting. The singular, unique feature of CoV-2 is that it requires a host enzyme called furin to activate it at a spot called the S1/S2 junction. No other coronavirus in the same subgenera has a furin cleavage site, as it is called. The other coronaviruses are cleaved at a site downstream from the S1/S2 site, called the S' site.

This is of course a major problem for the zoonosis theory, but it gets worse.

Since 1992 the virology community has known that the one sure way to make a virus deadlier is to give it a furin cleavage site at the S1/S2 junction in the laboratory. At least eleven gain-of-function experiments, adding a furin site to make a virus more infective, are published in the open literature, including Dr. Zhengli Shi, head of coronavirus research at the WIV. This has caused a flurry of Chinese papers since the pandemic began trying to show a natural furin site in a related virus (this one example was later shown to be an error in interpretation) or to show that furin sites from distant cousins of CoV-2 might be the source through a process called recombination, where two different viruses infect the same host and then make a mistake in copying their genetic material, and swap sequences.

These convoluted, hypothetical methods each fail, however. It turns out that it is Daszak himself who has shown that the subgenera of coronaviruses that have furin sites are found in different bat hosts, which live in different regions of China, than the sarbecovirus subgenera of which CoV-2 is a member. And even with these barriers, they apparently are too far apart to recombine. "For the three focal subgenera, *Sarbecoviruses, Merbecoviruses and Embevoviruses*…none of the three focal subgenera recombines with one another."[7] As noted previously[2] Dr. Shi also does not believe the bats of Hubei province are capable of being a host for CoV-2-related coronaviruses.

But it gets worse still for the zoonosis theory. The gene sequence for the amino acids in the furin site in CoV-2 uses a very rare set of two codons, three letter words so six letters in a row, that are rarely used individually and have never been seen together in tandem in any coronaviruses in nature. But these same 'rare in nature' codons turn out to be the very ones that are always used by scientists in the laboratory when researchers want to add the amino acid arginine, the ones that are found in the furin site. When scientists add a dimer of arginine codons to a coronavirus, they invariably use the word, CGG-CGG, but coronaviruses in nature rarely (<1%) use this codon pair. For example, in the 580,000 codons of 58 Sarbecoviruses the only CGG pair is CoV-2; none of the other 57 sarbecoviruses have such a pair.[8]

[7] CoV-2 is in the subgenera Sarbecoviruses.
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009272
[8] https://virological.org/t/alignment-of-58-sarbecovirus-genomes-for-conservation-analysis-of-sars-cov-2/430

So, there is no natural example of a furin protein site in nature that could be introduced into CoV-2 by recombination, there is no natural example of the particular gene sequence for the furin protein site contained in CoV-2 being used to code for anything in nature, but this particular coding is exactly what Dr. Shi, Baric, and others have used previously in published experiments to insert or optimize arginine codons.

It is telling that when Dr. Shi introduced the world to CoV-2 for the first time in January 2020 she showed hundreds of gene sequences of this novel virus but stopped just short of showing the furin site, the one she is purported to have introduced, seemingly not wanting to call attention to her handywork. She apparently failed to realize that an accomplished but innocent virologist, finding the first furin site ever seen in this class of viruses apparently coming from nature, would have featured the presence of the furin site prominently, and also would have used its presence and her experience with furin sites in other viruses to predict what it would foretell for the world due to its aggressive nature.

She could have perhaps saved many lives just by telling the world that she saw a furin site in the virus sequence. It would be left to a French and Canadian team to later identify the furin site in a paper.[9] They would write: "This furin-like cleavage site…may provide a **gain-of-function** to the 2019-nCoV for efficient spreading in the human population compared to other lineage b betacoronaviruses." [Emphasis added.]

Dr. Shi has denied the virus came from her lab, but she has created such a record of multiple examples of obfuscation, half-truths, contrived specimens, genetic sequences taken from thin air but published in premier journals and US NIH databases, etc. that her veracity is deeply damaged. Perhaps her words and actions on December 30, 2019 show the truth. Her very first response when told there was an unknown outbreak in Wuhan and to return back quickly from a meeting she was attending in Shanghai was to say, **"Could this have come from our lab?"**[10]

"I wondered if [the municipal health authority] got it wrong," she says. "I had never expected this kind of thing to happen in Wuhan, in central China." Her studies had shown that the southern, subtropical provinces of Guangdong, Guangxi and Yunnan have the greatest risk of coronaviruses jumping to humans from animals—particularly bats, a known reservoir. After all, the US equivalent of the distance, climate change, and human population density change between Yunnan and Wuhan is comparing the Everglades National Park in Florida and New York City.

Her other action on December 30 was to alter WIV computer databases of novel coronaviruses used by the world's virologists for research to make it more difficult to search for which coronaviruses she had in her building. In short, the day she was asked to address the pandemic in Wuhan, she chose to spend time to make unavailable to her fellow scientists of the world her decades of coronavirus work.

[9] https://www.sciencedirect.com/science/article/pii/S0166354220300528?via%3Dihub
[10] https://www.scientificamerican.com/index.cfm/_api/render/file/?method=inline&fileID=E1FDF8DE-9E22-4CE5-AD8B2E4682F52A86

The notion that CoV-2 was a laboratory creation, designed for maximum virulence, that escaped the laboratory accidentally has additional rings of evidence. From President Xi announcing in February new laws about laboratory security, to abundant evidence that the WIV was closed in October with few personnel inside, to the top military medical research doctor, General Chen Wei, being placed in charge of the WIV, to many more clues, it is clear an event occurred in Wuhan sometime in late 2019 that is most consistent with a laboratory escape.

The Asian region has a two-decade record of a little less than one laboratory-acquired infection per year. After the first SARS-CoV-1 epidemic was ended, SARS-CoV-1 jumped four more times into the human population, all from laboratories, with two in China. The last smallpox death in the entire world was a secretary who worked two floors above a research lab in England and contracted it through the ventilation system. The head of that laboratory committed suicide over his anguish for causing her death.

Over and over again. there is a long history and record of laboratory acquired infections that provides the background for considering what happened here.

Lab-made Bio-Weapon Hypothesis

But was SARS-CoV-2 more than just a gain-of-function experiment that escaped a laboratory? Could it have been one part of a two-part novel virus-vaccine bioweapons program?

General Chen Wei has been involved in vaccine research since joining the People's Liberation Army after college. In a 2017 internal speech at the AMMS (Academy of Military Medical Sciences) she said: *"只要有矛, 才能研究盾."* which translates roughly as, "you need to have an arrow to study a shield." I believe a Rubicon has been crossed by the world with this pandemic and framing the proper understanding of how we got here, and the proper response will be the critical next steps.

When Oppenheimer saw the application of Einstein's physics in the embodiment of the atomic bomb, he is said to have quoted a line from the Hindu scripture, the Bhagavad Gita, which reads: 'Now I am become Death, the destroyer of worlds.' The contribution of physics' research to human killing would total less than 300,000 people in two ten-square mile zones in Japan, and the horrors of those events led the world to regulate the raw materials of such bombs and to sanction sovereign nations who attempted to violate the rules.

This had followed the contribution of chemistry to human killing in the form of chemical warfare during World War I, in which 100,000 were killed, and led the nations of the world to an historic agreement to never use chemical warfare again. It is now only 'rogue' operators who violate the norms civilized nations have agreed to.

It seems to be biology's turn to show its dark arts. If it is generally understood that biology/biotechnology has been harnessed to create a pandemic that has killed more people than physics and chemistry research combined, and to be a weapon where no place on earth is safe from its effects (SARS-CoV-2 has been detected in the deepest Amazon jungles and at research stations in Antarctica), there needs to be developed a new set of regulations, rules, etc. to both honor the 1.8 million innocent people who died from COVID-19 and to protect the world so this

never happens again. It is also urgent to gather further data to support or refute if this was a Chinese bioweapons program, as the consequences of that would be significant.

Pre-publication peer review. The manuscript was provided by email to the following medical and scientific peers to afford an opportunity to review, comment, and critique the manuscript before publication. Those highlighted in yellow are members of the WHO-convened Global Study of the Origins of SARS-CoV-2[11], The Lancet COVID-19 Commission[12], or both.

First Name	Last Name
John	Amuasi
Kristian	Andersen
Danielle	Anderson
Ralph	Baric
Francis	Collins
Carlos	das Neves
Peter	Daszak
Vladimir	Dedkov
Dominic	Dwyer
Anthony	Fauci
Hume	Field
Tedros Adhanom	Ghebreyesus
Eddie	Holmes
Gerald	Keusch
Marion	Koopmans
Dato' Sai Kit (Ken)	Lam
Fabian	Lendertz
W. Ian	Lipkin
Ken	Maeda
Hung	Nguyen
Stanley	Perlman
David	Quammen
Andrew	Rambaut
Angelie	Rassmussen
Linda	Saif
Zhengli	Shi
Supaporn	Wacharapluesadde

[11] https://www.who.int/health-topics/coronavirus/origins-of-the-virus
[12] https://covid19commission.org/origins-of-the-pandemic

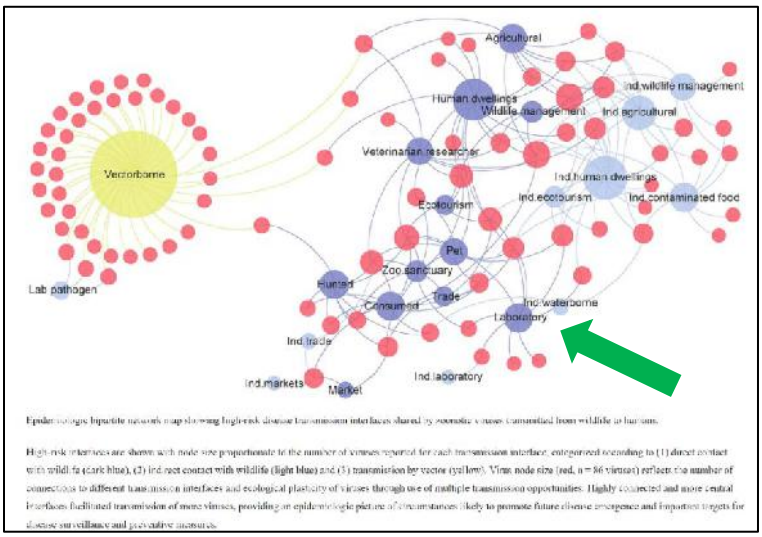
A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

Introduction. A two-hypothesis, Bayesian analysis was conducted to determine the origin of the SARS-CoV-2 pandemic. The conclusion was that it was created in a laboratory with synthetic biology tools from a bat beta coronavirus, subgenera sarbecovirus backbone (98.9% probability) and not from a natural, zoonotic transmission (1.1%).

There is no direct evidence of whether the release was accidental, or deliberate but circumstantial evidence makes it is highly likely it was accidental.

At the one-year anniversary of the first cases of COVID-19, the coronavirus pandemic caused by the SARS-CoV-2 virus, the origin of the virus remains unknown. While leading institutions and experts have been consistently adamant that it is a zoonotic disease which jumped from a bat reservoir host to humans directly or through an intermediate host the alternative possibility that it escaped from a laboratory conducting research remains a viable option.

In fact, in 2015 Peter Daszak, a leading zoonotic proponent of CoV-2 origin, wrote in, "Spillover and pandemic properties of zoonotic viruses with high host plasticity,"[13] that transmission from laboratories was a major source of zoonotic disease. The Figure below from the Daszak paper shows this important relationship (green arrow):

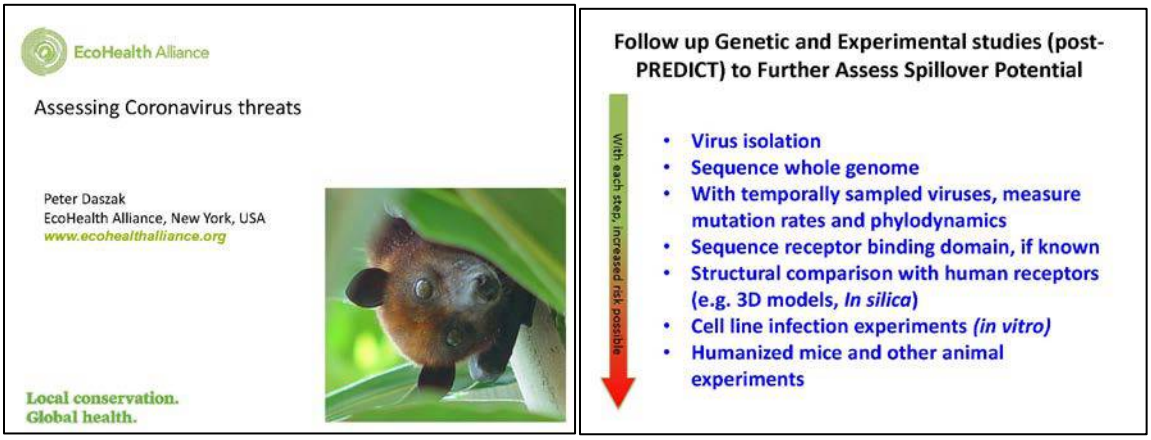


Daszak et al. also writes: "**Zoonotic virus spillover** from wildlife was most frequent in and around human dwellings and in agricultural fields, as well as **at interfaces with occupational exposure to animals** (hunters, **laboratory workers**, veterinarians, **researchers**, wildlife management, zoo and sanctuary staff). **Primate hosts were most frequently cited as the source of viruses transmitted by direct contact** during hunting (exact P = 0.051) and **in laboratories**

[13] https://www.nature.com/articles/srep14830

(exact P = 0.009)." [Emphasis added]. Primate "hosts" can presumably include monkey cell culture, such as the ubiquitous VERO cell used in all virology laboratories, including the WIV.

In 2015 Dr. Daszak spoke of the spillover danger of certain types of laboratory research:



He writes: "with each step, increased risk possible" with "Humanized mice and other animal experiments" the highest risk work.

In a prescient Twitter post in November 2019, he highlights the work he is doing using recombinant viruses with humanized mice and making viruses that "**don't respond to MAbs, vaccines…**" in response to criticism his work is of limited value:

Clearly, before the beginning of the pandemic, Daszak, now a member of both the WHO and Lancet teams being sent to China to explore the origin of CoV-2, could entertain the eal possibility of a laboratory created virus escaping into the human population/community.

The purpose of this analysis is to use a Bayesian Inference Network approach to the collected circumstantial evidence that is available to provide likelihoods of the alternative hypotheses as to the origin of SARS-CoV-2. The analysis also will include certain prior probabilistic conclusions to help set the initial state before the proprietary evidence is used.

Origin hypotheses: Initial States to establish the posterior probabilities.

Two published Bayesian analyses and two independent studies of zoonotic spillover from nature and laboratory-acquired infections in Asia will be used to establish the posterior probabilities for this analysis.

Zoonotic spillover frequency versus laboratory acquired infection frequency based on two published papers, one by Daszak et al.

In 2015 Daszak et al. published a paper entitled, "Spillover and pandemic properties of zoonotic viruses with high host plasticity,"[1] in which they identified 162 zoonotic viruses with naturally occurring animal-to-human transmission from 1990-2010. This is a frequency of 162/20 = 8.1 events per year.

They also note: "The majority (94%) of zoonotic viruses described to date (n = 162) are RNA viruses, which is 28 times higher (95% CI 13.9–62.5, exact $P < 0.001$) than the proportion of RNA viruses among all vertebrate viruses recognized, indicating that RNA viruses are far more likely to be zoonotic than DNA viruses." CoV-2 is an RNA virus.

Finally, they note that: "In general, wild animals were suggested as the source of zoonotic transmission for 91% (86/95) of zoonotic viruses compared to 34% (32/95) of viruses transmitted from domestic animals and 25% (24/95) with transmission described from both wild and domestic animals."

One of the caveats of the Daszak data is that it categorizes a laboratory-acquired infection (LAI) from an animal collected from the wild as a zoonotic spillover. There is no data in the paper to assess this issue and leaving it uncorrected is a conservative approach since it only inflates the natural zoonotic frequency.

In 2018 a paper by Siengsanan-Lamont entitled, "A Review of Laboratory-Acquired Infections in the Asia-Pacific: Understanding Risk and the Need for Improved Biosafety for Veterinary and Zoonotic Diseases," was published.[14] They reported 27 LAIs between 1982 and 2016, a frequency of 27/(2016 – 1982) = 0.8 events per year.

Using these historical frequencies of zoonotic spillover versus LAI to predict a future event can be calculated in the following manner:

Evidence	Zoonotic Origin	Laboratory Origin
Frequency per year from Daszak paper	8.1	NA
Frequency per year from Siengsanan-Lamont paper	NA	0.8
Total events per year	8.1 + 0.8 = 8.9	8.1 + 0.8 = 8.9
Likelihood of future event based on historical frequency	8.1/8.9 X 100 = 0.91	0.8/8.9 X 100 = 0.9

Daszak's initial state analysis. This evidence sets the likelihood that CoV-2 was a zoonotic origin event at 91% and a laboratory origin event at 9%.

[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6073996/

Independent prior analyses: Rootclaim.

The next data that will be used is a recent analysis published on the Rootclaim website.[15] Three hypotheses below were analyzed through a series of evidence statements and the probabilities that each was the origin of SARS-CoV-2 determined:

Hypothesis	Calculated Probability
Lab escape: The virus was the subject of genetic research, including gain-of-function, and was released by accident	81%
Zoonotic: The virus evolved in nature and was transmitted to humans from a non-human vertebrate animal	16%
Bioweapon: The virus was genetically engineered as a bioweapon and was deliberately released	3%

As can be seen, the highest likelihood probability is an accidental lab escape, the lowest a bioweapon. The details of the evidence used to arrive at this conclusion is contained in Appendix 1. A summary of the changes in probability at each level of evidence analysis is shown in this table:

Evidence	Laboratory	Zoonosis	Bioweapon
Starting point	1.2%	82%	16%
Contagion and mortality	1.4%	97%	1.9%
Outbreak location: Wuhan	42%	56%	2.8%
Virus sources near Wuhan	16%	83%	1.0%
Chimera	37%	60%	2.5%
Furin cleavage	72%	23%	4.8%
WIV lab procedures	80%	17%	3.5%
WIV disassociation	89%	9%	2.0%
Chinese response	90%	8%	1.7%
No reported infections at WIV	86%	11%	2.4%
No whistleblowers	81%	16%	2.8%

As can be seen, the starting point assumed an 82% probability of a zoonotic origin. This starting point is a reasonable value and will be used here. Since some of the evidence in the above analysis will be used here, only the starting point will be used and not the probability changes from there.

For purposes of this analysis only the Rootclaim initial state will be used since much of their evidence is also covered in the analysis here.

[15] https://www.rootclaim.com/analysis/what-is-the-source-of-covid-19-sars-cov-2

In a paper by Daszak and colleagues it states: "In general, wild animals were suggested as the source of zoonotic transmission for 91% (86/95) of zoonotic viruses compared to 34% (32/95) of viruses transmitted from domestic animals and 25% (24/95) with transmission described from both wild and domestic animals."[1]

On the other hand, domestic animals seem to have been ruled out for SARS-CoV-2. In an interview for *Science* in July 2020, Dr. Zhengli Shi, head of coronavirus research at the Wuhan Institute of Virology, stated: "Under the deployment of the Hubei Provincial Government, our team and researchers from Huazhong Agricultural University collected samples of farmed animals and livestock from farms around Wuhan and in other places in Hubei Province. We did not detect any SARS-CoV-2 nucleic acids in these samples."[16]

Reanalysis of Rootclaim initial state to remove Bioweapons option.

The US government uses the following definitions:

"Gain-of-function (GOF) studies, or research that improves the ability of a pathogen to cause disease, help define the fundamental nature of human-pathogen interactions, thereby enabling assessment of the pandemic potential of emerging infectious agents, informing public health and preparedness efforts, and furthering medical countermeasure development.

Gain-of-function studies may entail biosafety and biosecurity risks; therefore, the risks and benefits of gain-of function research must be evaluated, both in the context of recent U.S. biosafety incidents and to keep pace with new technological developments, in order to determine which types of studies should go forward and under what conditions."[17]

"Dual use research of concern (DURC) is life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops, and other plants, animals, the environment, materiel, or national security. "[18]

For this analysis, the assumption is made that GOF and DURC are largely the same processes and techniques in the laboratory and thus can only be distinguished by direct, documentary evidence of the intent of the research from administers in the facilities conducting the work.

In the absence of any such documentary evidence that bioweapon research was being conducted or that SARS-CoV-2 is a bioweapon and to take the least inflammatory posture, the initial state for the above prior analysis will be recalculated by eliminating the hypothesis, and its accompanying probability, that SARS-CoV-2 was created as a bioweapon. The revised initial state calculation is shown in this table:[19]

[16] https://www.sciencemag.org/sites/default/files/Shi%20Zhengli%20Q%26A.pdf
[17] https://www.phe.gov/s3/dualuse/Pages/GainOfFunction.aspx
[18] https://www.phe.gov/s3/dualuse/Pages/default.aspx
[19] For clarity, the 3% bioweapon probability was simply dropped and the remaining likelihoods, 81% and 16%, were normalized.

Evidence	Zoonotic Origin	Laboratory Origin	Bioweapons Origin
Rootclaim initial state	0.86	0.012	0.16
Remove bioweapons	NA	NA	0
Normalize remaining hypotheses	0.86/(0.86 + 0.012) = 0.986	0.012/(0.86 + 0.012) = 0.014	NA

Rootclaim Initial state analysis, adjusted. This evidence sets the likelihood that CoV-2 was a zoonotic origin event at 98.6% and a laboratory origin event at 1.4%.

Additional Prior Evidence by Demaneuf and De Maistre. A second prior Bayesian analysis was performed by professionally educated risk assessment personnel and Chinese-language speaking professionals[20] and is included herein in its entirety. For the sake of brevity, the zoonotic origin evidence was based primarily on population size, distribution, and geographic distribution of bat populations relative to Wuhan. With respect to a lab accident, they separately analyze probabilities of a virus escape during collection, transport, and direct lab accidents and then separately the probability of a community outbreak following a lab escape. They also use primary Mandarin-language sources for Chinese estimates of the same events, showing corroboration of the probabilities. Their conclusion is that the probability of a lab escape ranges from 6% to 55% with a zoonotic origin a zoonotic origin probability being 45% to 94%.

Second Bayesian analysis. Using the most conservative probabilities, this evidence sets the likelihood that CoV-2 was a zoonotic origin event at 94% and a laboratory origin event at 6%.

Selection of initial state for Bayesian analysis.

The Text-Table below summarizes the three approaches to an initial state as to the origin of CoV-2. While the Demaneuf and De Maistre analyses set a range for the zoonotic origin of 45% to 94%, I have used the top of the range of their probability of a zoonotic origin to be conservative.

Prior Analysis	Zoonotic Origin	Laboratory Origin
Daszak et al. paper	91%	9%
Rootclaim Bayesian analysis	98.6%	1.4%
Demaneuf and De Maistre Bayesian analysis	94%	6%

Using a simple online calculator[21] the mean of these three value sets is 94.5%, the standard deviation is $\pm$ 3.8%, and the 95% confidence interval is $\pm$ 4.3%. Using these data, the upper bound of the 95% confidence interval is 98.8% and, to be most conservative, this will be used as the starting probability of a zoonotic origin.

Initial state for this analysis. The likelihood that SARS-CoV-2 began as a zoonotic event is 98.8% and the likelihood it began as a laboratory event is 1.2%.

[20] https://zenodo.org/record/4067919#.X-qIm9gzbOj . For reference purposes, this paper comes with a spreadsheet listing 112 individual BSL-3 labs in China across 62 lab-complexes.
[21] https://www.calculator.net/standard-deviation-calculator.html?numberinputs=91%2C+94%2C+98.6&ctype=s&x=48&y=19

1. General approach of this analysis[22]

This analysis is intended to examine two competing and mutually exclusive theories of the origin of the coronavirus, SARS-CoV-2 (CoV-2), and the pandemic it has caused, COVID-19.

At the time of this writing there have been 83 million confirmed cases and 1.8 million deaths.[23] Some sources place the economic damage at $21 trillion USD.

Bayes Theorem

This brief description of the Bayes Theorem was taken from the work of Jon Seymour:[24]

"The eponymously named Bayes Theorem was discovered by the Reverend Thomas Bayes in the 1700's and saved for posteriority by an archivist of his papers who discovered the work posthumously. In common language, it provides a rational technique for revising a prior belief in light of new evidence. The equation for Bayes Theorem is given below:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

where:

- H is the statement of the hypothesis of interest
- P(H) is the prior probability that the hypothesis is true, independent of the evidence.
- E is the evidence being used to revise the belief in hypothesis
- P(E) is the marginal likelihood of the evidence, independent of the hypothesis
- P(E|H) is the likelihood the evidence, given that the hypothesis is true
- P(H|E) is the posterior probability of the hypothesis, given the evidence.

P(E) is sometimes difficult to estimate, but the following identity must hold:

$$P(E) = P(E|H).P(H) + P(E|\hat{H}).P(\hat{H})$$

Here P(E|^H) is the probability of the evidence, assuming the hypothesis is false and P(^H) is the probability the hypothesis is false which is the same as 1-P(H). Estimating the two conditional probabilities P(E|H) and P(E|^H) is generally easier than estimating the unconditional probability, P(E)."

---

[22] The statistical approach and many of the individual statistical analyses were performed by Dr. Martin Lee, PhD, Adjunct Professor of Biostatistics, UCLA. https://ph.ucla.edu/faculty/lee The likelihood adjustments to the Bayesian analysis, which you can see are routine math, were conducted by the author.
[23] https://www.worldometers.info/coronavirus/coronavirus-cases/
[24] https://jonseymour.medium.com/a-bayesian-analysis-of-one-aspect-of-the-sars-cov-2-origin-story-where-the-first-recorded-1fbdcbea0a2b

Theory One. The zoonotic theory is that a vertebrate animal was infected with CoV-2 or an ancestor (Index Host) and that a human was infected with contact to that Index Host in some manner. Human-to-human spread then followed.

Theory Two. The laboratory origin theory is that CoV-2 or an ancestor was being used in laboratory experiments and that it 'escaped' from the lab via an infected person, lab animal, experimental waste, etc.

I have found no evidence of a deliberate release and early firsthand accounts of local officials and scientists suggest surprise and consternation. If this was a deliberate release, such evidence would be extremely local, limited in distribution, and highly compartmentalized. It is beyond the scope of this analysis.

Weight of the evidence. For purposes of the calculation of posterior probabilities in the Bayesian analysis, evidence which has a statistical basis will be used directly to adjust the probabilities.

Statistically significant evidence. Since some of the probability calculations have astronomical values which would make a single such evidence statement, if inputted directly, swamp any further calculation and make their later contribution mute, a decision was made to simply treat quantitative probabilities as significant at the $p = 0.05$ level, no matter how much 'more significant' the calculation suggested.

So, for example, a probability of certain codon usage coming from nature may be one in 440 or $p = 0.002$, the contribution of this evidence to the input to the posterior probability adjustment would be set at a p-value of 0.05. In such cases the adjustment would be to change the 'winning' hypothesis by multiplying by 19, since a $p = 0.05$ is the same as a 19 out of 20 likelihood event. This is a conservative treatment of what would be highly significant data.

Other quantitative evidence. If a piece of evidence can be quantified but it does not reach a significance of $p = 0.05$ it will be used directly in the likelihood adjustment.

Non-quantitative evidence. For evidence that cannot be quantified, the decision was made to treat these as quantitative outcomes with a 51% to 49% likelihood value with respect to the 'winning' hypothesis. This has the effect of increasing the probability of that hypothesis for that step in the Bayesian analysis by 1.04. This 51%/49% concept is related to the legal standard of the 'preponderance of the evidence' used in civil litigation.

Independence. An important qualitative assessment that must be made is whether or not two pieces of evidence are independent of each other. If they are independent, they can each be used in determining a new likelihood calculation. If they are dependent on each other then they must be combined and only a single new likelihood analysis can be made. Where ever possible, evidence statements that could be considered as dependent are called out and this rule is followed on their contribution to the analysis.

Subjective Discount Factor. The impact of each piece of evidence was adjusted further by a subjective discount factor. This is a qualitative assessment of the overall veracity of a particular

piece of evidence when all factors, samples, methods, data sources, etc. are taken into context. It varies from 60% to 100% and is used as a fraction to reduce the impact of a single piece of evidence even further.

Hearsay. Just as in a court of law, evidence, usually attributed to a given person or persons, that is not directly available but instead relies on statements of others is usually not allowed in a court trial and will accordingly not be used here to adjust the Bayesian analysis. It may be recorded and preserved as a placeholder and reminder for further research. If new, direct evidence can be found than the bar of using it is lifted and it can be used for adjustment.

Significant figures. Because of the overall nature of the analyses here, all math calculations related to likelihoods are performed and carried forward at the 'one significant figure' level, with standard rounding rules applied. This has the effect, near the end of the cumulative evidence, of failing to change the relative probabilities as the small adjustments are reversed in the rounding process.

Evidence. International committees to investigate the origin of SARS-CoV-2 may not be impartial.

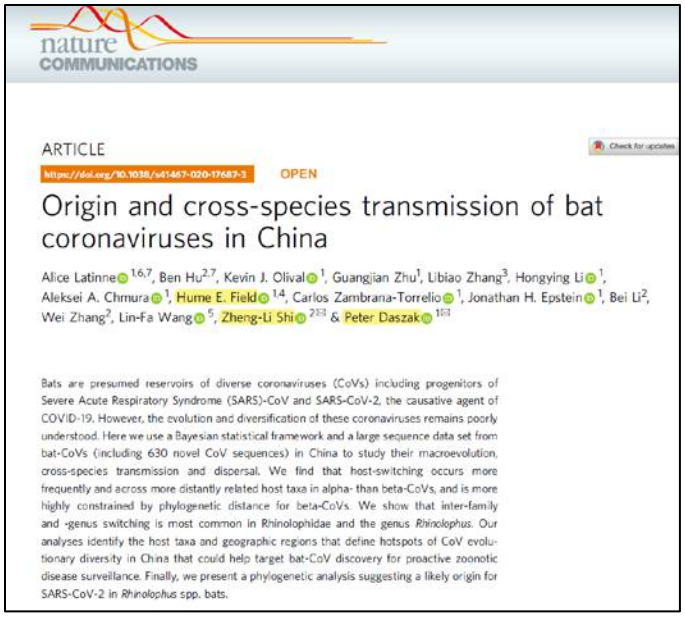
At the time of the writing of this manuscript there are two committees charged with examining the evidence and determining the origin of the SARS-CoV-2 virus. One committee is commissioned by the World Health Organization (WHO) and the other is an ad hoc committee established by the British medical journal, *The Lancet.*

The composition of the two committees is shown in the Text-Table below:

	Lancet Commission of CoV-2	WHO Commission on CoV-2 origin
	Dr. Peter Daszak, Chair	**Dr. Peter Daszak, Ph.D (EcoHealth Alliance, USA)**
	Dr. John Amuasi	Prof. John Watson (Public Health England, United Kingdom)
	Dr. Danielle Anderson	Prof. Dr. Marion Koopmans, DVM PhD (Erasmus MC, Netherlands)
	Dr. Isabella Eckerle	Prof. Dr. Dominic Dwyer, MD (Westmead Hospital, Australia)
Also co-author	Dr. Hume Field	Vladimir Dedkov, Ph.D (Institute Pasteur, Russia)
	Dr. Gerald Keusch	Dr. Hung Nguyen, PhD (International Livestock Research Institute (ILRI), Vietnam)
	Dr. Dato' Sai Kit (Ken) Lam	PD. Dr. med vet. Fabian Lendertz (Robert Koch-Institute, Germany)
	Dr. Carlos das Neves	Prof. Dr. Thea Fisher, MD, DMSc(PhD) (Nordsjællands Hospital, Denmark)
	Dr. Malik Peiris	Dr. Farag El Moubasher, Ph.D (Ministry of Public Health, Qatar)
	Dr. Stanley Perlman	Prof. Dr. Ken Maeda, PhD, DVM (National Institute of Infectious Diseases, Japan)
	Dr. Linda J. Saif	WHO Commission of CoV-2 origin
	Dr. Supaporn Wacharapluesadee	
	Lancet Commission on CoV-2	
	Signed Lancet letter	
	Co-author with Daszak	

There are a number of potential conflicts of interest:

Fully half of The Lancet's team had already suggested that any lab-leak hypothesis was a "conspiracy theory" in a January 2020 paper that has been shown elsewhere within to have been orchestrated behind the scenes to appear spontaneous.



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-17687-3 OPEN

Origin and cross-species transmission of bat coronaviruses in China

Alice Latinne[1,6,7], Ben Hu[2,7], Kevin J. Olival[1], Guangjian Zhu[1], Libiao Zhang[3], Hongying Li[1], Aleksei A. Chmura[1], Hume E. Field[1,4], Carlos Zambrana-Torrelio[1], Jonathan H. Epstein[1], Bei Li[2], Wei Zhang[2], Lin-Fa Wang[5], Zheng-Li Shi[2✉] & Peter Daszak[1✉]

Bats are presumed reservoirs of diverse coronaviruses (CoVs) including progenitors of Severe Acute Respiratory Syndrome (SARS)-CoV and SARS-CoV-2, the causative agent of COVID-19. However, the evolution and diversification of these coronaviruses remains poorly understood. Here we use a Bayesian statistical framework and a large sequence data set from bat-CoVs (including 630 novel CoV sequences) in China to study their macroevolution, cross-species transmission and dispersal. We find that host-switching occurs more frequently and across more distantly related host taxa in alpha- than beta-CoVs, and is more highly constrained by phylogenetic distance for beta-CoVs. We show that inter-family and -genus switching is most common in Rhinolophidae and the genus *Rhinolophus*. Our analyses identify the host taxa and geographic regions that define hotspots of CoV evolutionary diversity in China that could help target bat-CoV discovery for proactive zoonotic disease surveillance. Finally, we present a phylogenetic analysis suggesting a likely origin for SARS-CoV-2 in *Rhinolophus* spp. bats.

The above paper published in August 2020 has as co-authors Drs. Hume, Daszak, and Shi. Having two of these scientists be asked to investigate a third co-author is a clear conflict of interest.

A newspaper piece about Peter Daszak entitled, "The doctor who denied COVID-19 was leaked from a lab had this major bias,"[25] questions his ability to be unbiased due to a deep, long history of work with Dr. Zhengli Shi of the WIV.

A lengthy piece in Wired was subtitled, "The two major investigations into the origins of the pandemic are compromised by potential conflicts of interest."[26]

Since the purpose of this manuscript is to evaluate the scientific evidence concerning the origin of SARS-CoV-2 no further effort will be put into these matters. If and when a report is prepared from either committee there will be time to analysis the work in the reports and compare it to prior publications and statements from the committee members to look for bias.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

[25] https://nypost.com/2021/01/16/doctor-who-denied-covid-was-leaked-from-a-lab-had-this-major-bias/
[26] https://www.wired.com/story/if-covid-19-did-start-with-a-lab-leak-would-we-ever-know/?utm_source=twitter&utm_medium=social&utm_campaign=onsite-share&utm_brand=wired&utm_social-type=earned

Evidence. Three high visibility papers grounded the zoonotic origin hypothesis in the public conversation from February to May 2020: a pros and cons analysis.

Introduction. The two key data points from December 2019 concerning the origin of the SARS-CoV-2 coronavirus infection, the cause of COVID-19, are the observation that a large number of the earliest patients worked or had visited the Hunan Seafood Market in Wuhan, China and that the hospitals where the first patients were admitted were a short distance from the Wuhan Institute of Virology (WIV), the only high security, BSL-4 laboratory in all of China, and arguably the leading research institute in the world studying coronaviruses of the type causing COVID-19.

The first data point is reminiscent of the origin of SARS-CoV-1, a zoonosis with interspecies transmission from bats to civet cats and then to humans, identified in wet markets in southern China. The second data point is reminiscent of the four SARS-CoV-1 human spillovers that occurred after the 2003 epidemic ended and were each a laboratory-acquired infection (LAI) by a scientist working in a government research laboratory, much like the WIV, and then local human-to-human spread and nearby hospital admission.

To be clear in this paper, the term zoonosis will only be used to describe a interspecies transmission outside of a laboratory. This point seems important to clarify since Dr. Zhengli Shi, head of coronavirus research at the WIV, has previously reported: "An outbreak of hemorrhagic fever with renal syndrome occurred among students in a college (College A) in Kunming, Yunnan province, China in 2003. Subsequent investigations revealed the presence of hantavirus antibodies and antigens in laboratory rats at College A and two other institutions. Hantavirus antibodies were detected in 15 additional individuals other than the index case in these three locations. Epidemiologic data indicated that the human infections were a result of **zoonotic transmission** of the virus from laboratory rats."[27] [emphasis added.] The author has found no other support for the use of the term zoonotic transmission with respect to an LAI and its dual use could be confusing, and so will be avoided.

While the two initial data points would suggest that a balanced approach should be taken with respect to investigations of the origin of SARS-CoV-2, three high visibility publications that argued the laboratory origin idea was a "conspiracy theory" and strongly argued that it was of zoonotic origin foreclosed legitimate debate for much of 2019. The purpose of this evidence analysis is to examine these papers and weigh the strength of the evidence.

Paper 1: The February 3, 2020 paper by WIV scientist Dr. Shi et al. entitled: "A pneumonia outbreak associated with a new coronavirus of probable bat origin."

This seminal paper set the stage for the zoonotic origin of SARS-CoV-2 and has been accessed over one million times. According to *Nature*, this article is in the 99th percentile (ranked 24th) of the 326,159 tracked articles of a similar age in all journals and the 99th percentile (ranked 2nd) of the 783 tracked articles of a similar age in Nature.

[27] https://pubmed.ncbi.nlm.nih.gov/20380897/

However, a careful analysis of it shows serious issues which suggest it is unreliable. The following analysis is in the form of an independent manuscript:

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

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[28] 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[29]" 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

[28] 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 .

[29] 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."

samples 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.[30]

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.

[30] Dr. Shi Science interview July 2020

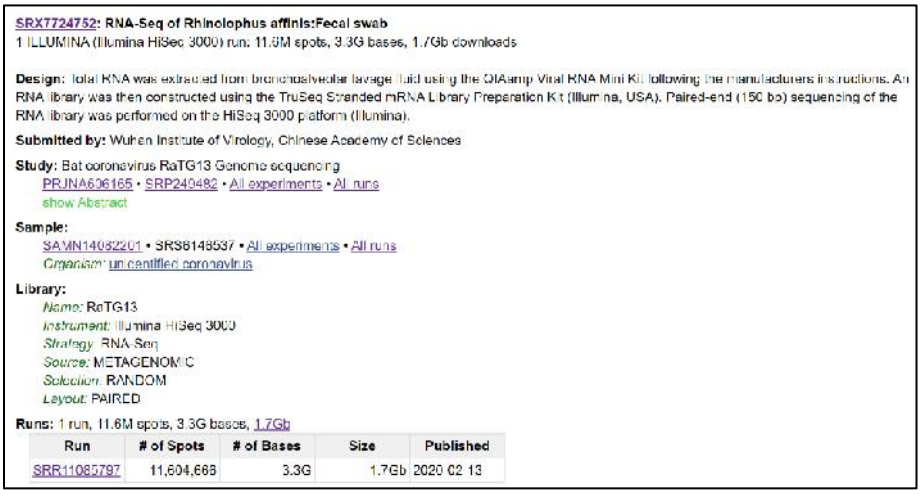
Materials and Methods.

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 Preperation 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 • SRS6148537 • All experiments • All runs
Organism: uncentified 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	11,604,668	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.[31] They have submitted three entries to GenBank for these two viruses, in 2016, February 2020, and May 2020.[32] 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.[33] 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, 34] Given this disclosure months after the original publication concerning RaTG13 in *Nature* it is possible that the omission of the original publication and sequence data concerning BtCoV/4991 violated the "Reporting

[31] 2016 Virologica Sinica paper and February 2020 Nature paper
[32] 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.
[33] The GISAID entry is EPI_ISL_402131.
[34] 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."

standards and availability of data, materials, code and protocols" required for *Nature* publications.[35]

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, we compared the sequence of SARS-CoV-2 and our unpublished bat

[35] Nature research reporting standards for availability of data

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.[36] Blast was used for sequence alignment and comparisons.[37]

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[38]" 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:

[36] NCBI Sequence Archive
[37] Blast alignment
[38] 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

> "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 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.

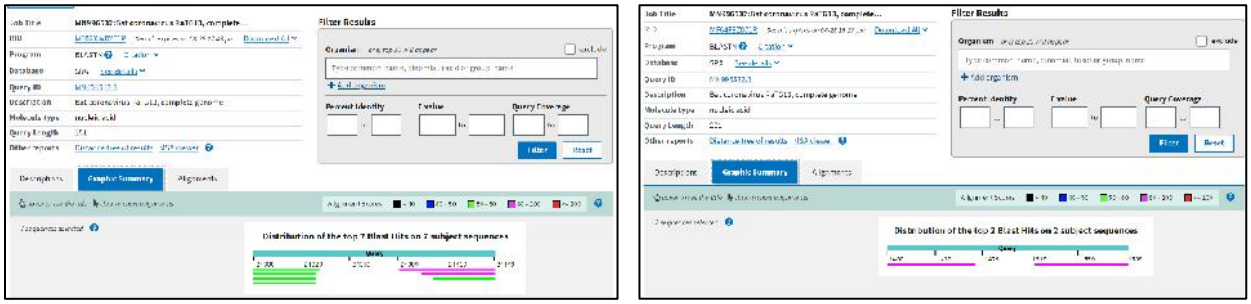
A Blast analysis of the RaTG13 genome against SRR11085797 retrieved about 1700 reads which covers only about 252,000 nt of the total reads of 3.3 Gb. Since the genome size of RaTG13 is known to be about 30,000 nt this represents an 8-fold coverage, typically insufficient for a definitive assembly. For example, some have suggested a 30-fold coverage is necessary to create high quality assemblies.[39]

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

At an eight-fold coverage and based on the typical practice of having four or more reads to call a SNP,[40] 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.

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:

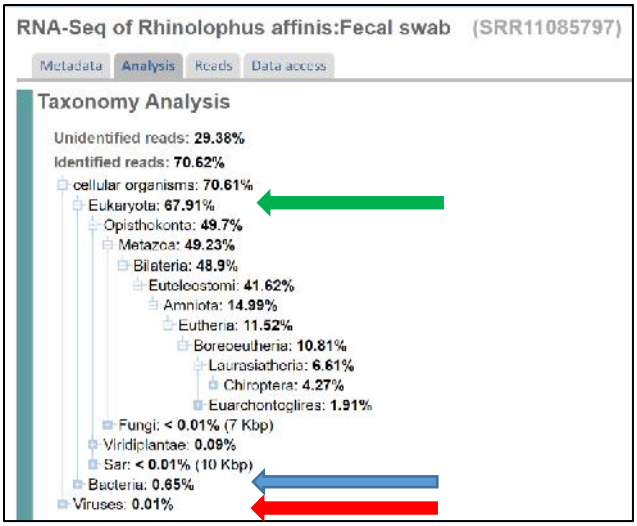


It is conceivable there are additional gaps but the above two are sufficient to document that the complete RaTG13 genome sequence could not have been assembled solely from the RNA-Seq data, as stated.[2]

Taxonomy analysis of the RaTG13 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%):

[40] Illumina Technical Bulletin Call Coverage



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,"[41] 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 RaTG13'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.

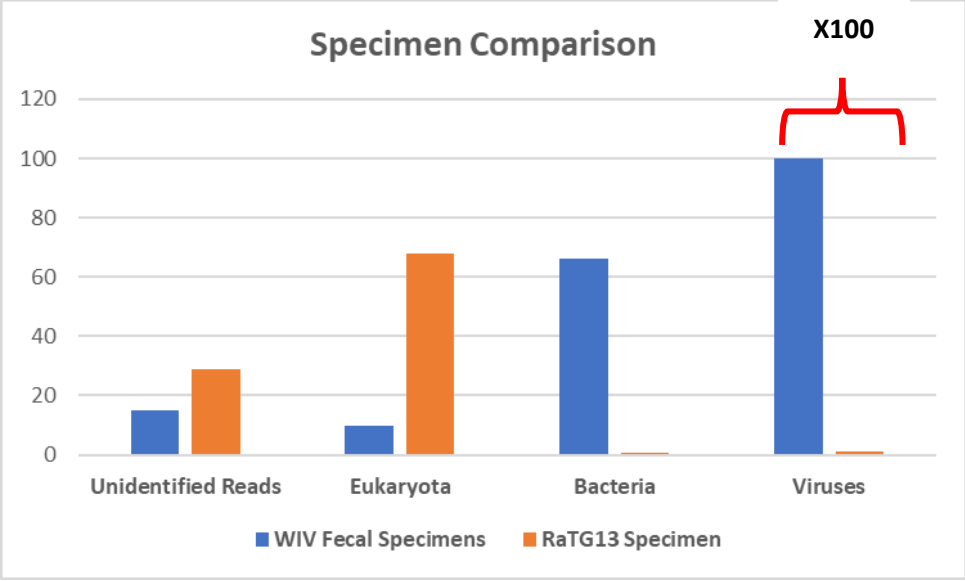
[41] Discovery of bat coronaviruses through surveillance and probe capture-based next-generation sequencing

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.

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.

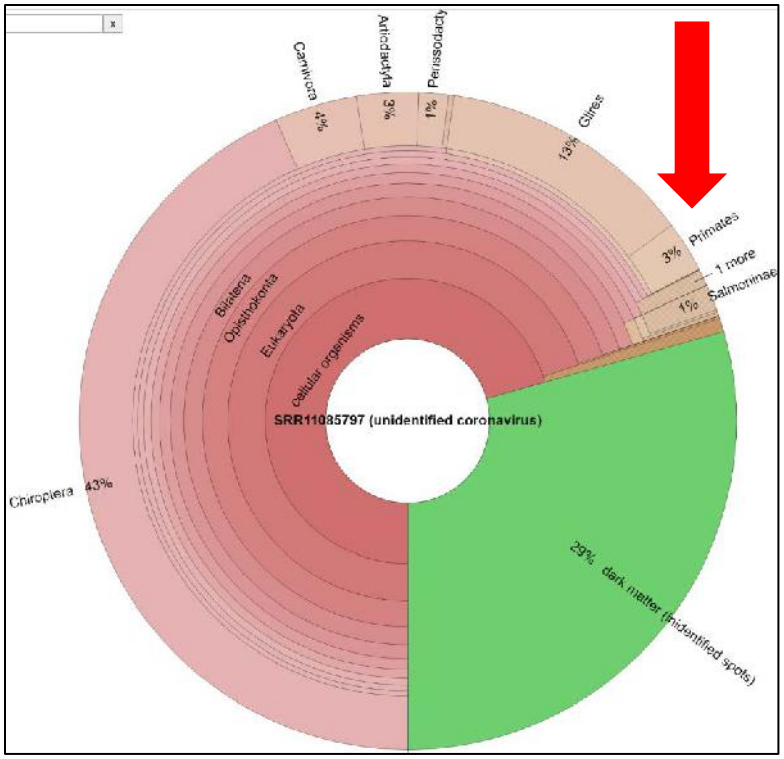
Specimen	The identity of the Strong Signals in the Specimens		
	Bacteria	Eukaryotes	Viruses
Rhinolophus affinis anal swab (SRR11085736)	92%	One magnaorder of placental mammals, includes bat	None
Miniopterus schreibersii anal swab (SRR11085734)	88%	**One bat**, the host bat, Miniopterus sp.	None
Scotophilus 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.



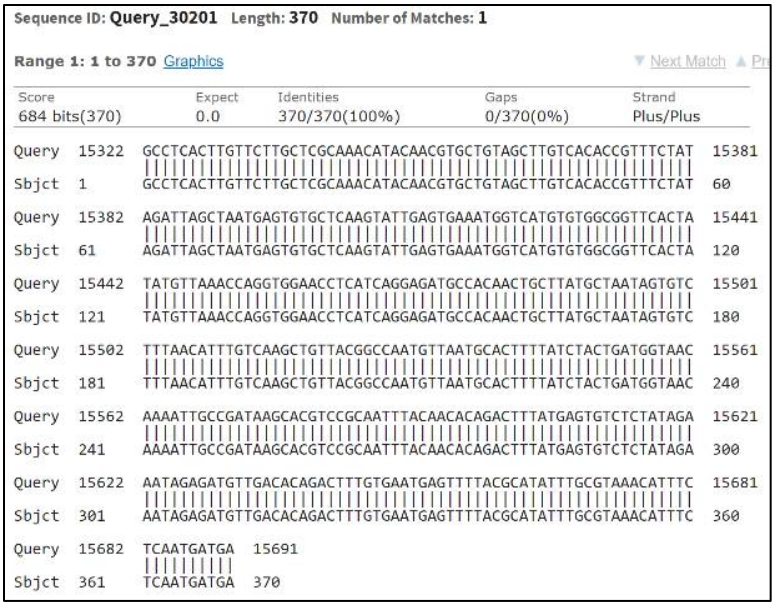
Source: [Krona analysis of RaTG13 specimen]

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.



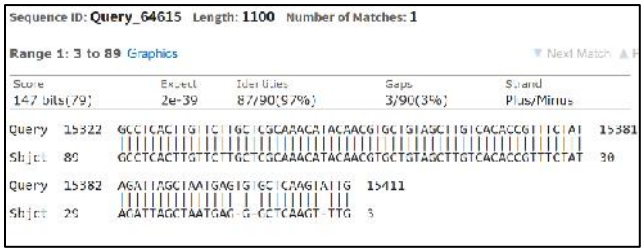
Sequence ID: **Query_30201** Length: **370** Number of Matches: **1**

Range 1: 1 to 370 Graphics

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).

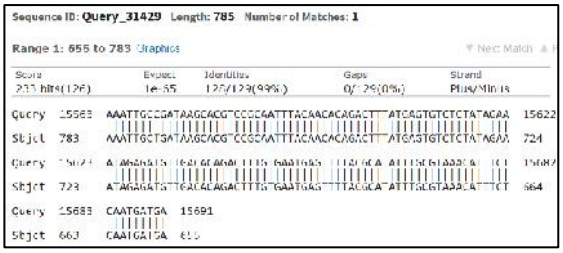


Sequence ID: **Query_64615** Length: **1100** Number of Matches: **1**

Range 1: 3 to 89 Graphics

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-CCTCAAGT-TTG  3
```



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

Range 1: 655 to 783 Graphics

Score	Expect	Identities	Gaps	Strand
235 bits(126)	1e-55	126/129(98%)	0/129(0%)	Plus/Minus

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

Query  15623  A-AAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  15681
              | |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  723    ATAAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  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).

Download Graphics SRA

SRX8357956

Sequence ID: **SRA:SRR11806578.14.1** Length: **1100** Number of Matches: **1**

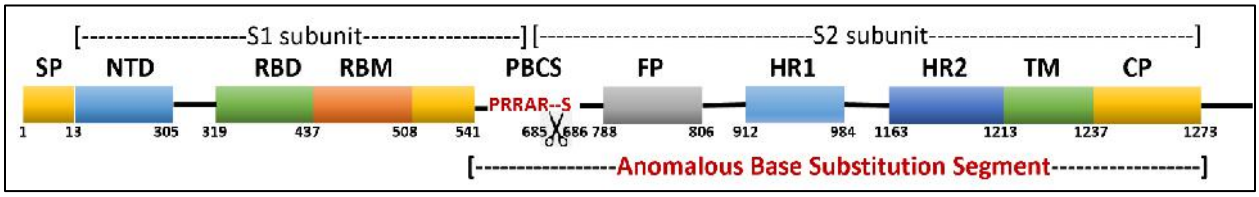
Range 1: 14 to 1100 Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
1716 bits(929)	0.0	1052/1107(95%)	25/1107(2%)	Plus/Minus

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:



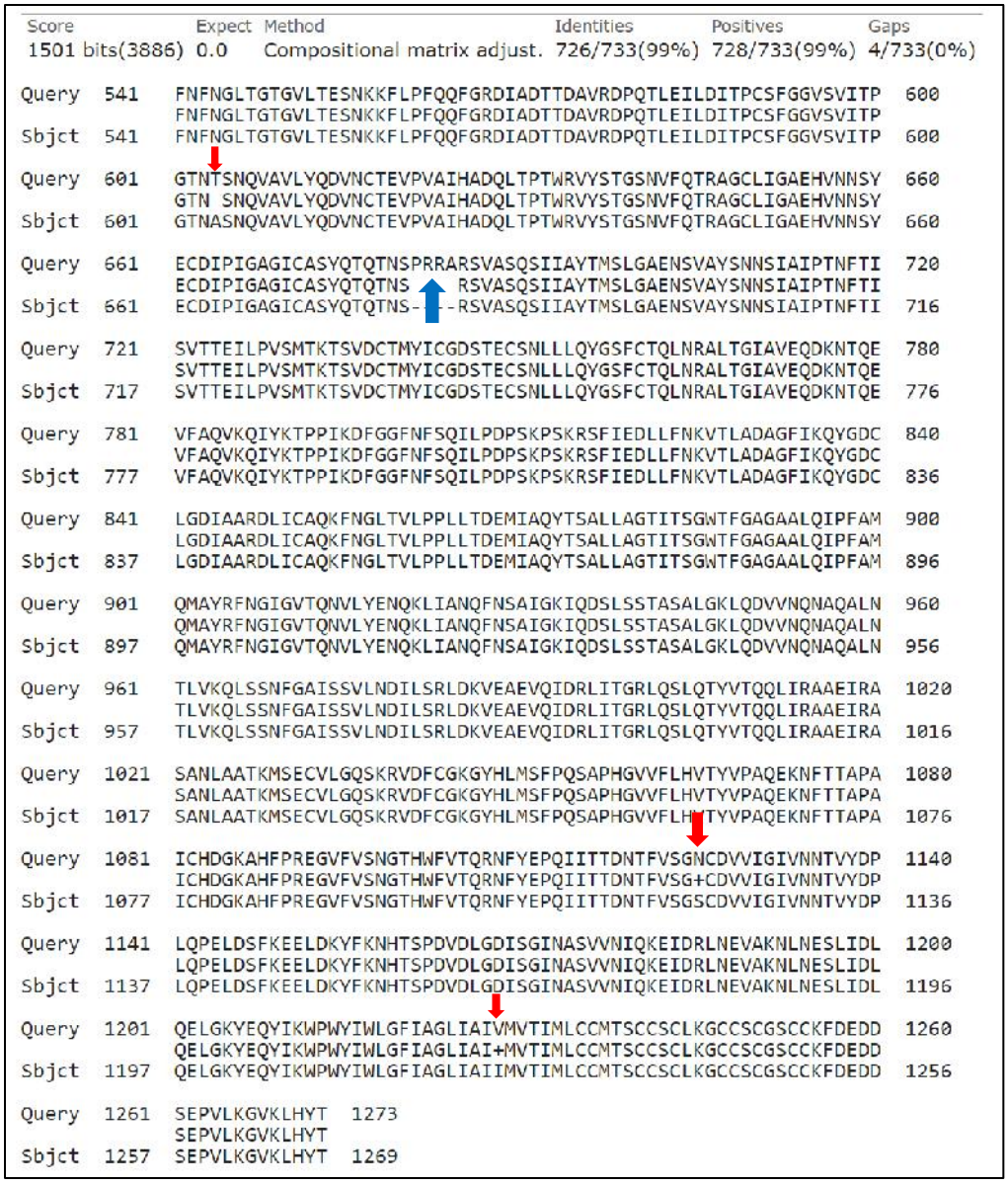
The SARS-CoV-2 Spike protein (above) contains an S1 subunit and S2 subunit with the Polybasic Cleavage Site (PBCS) between R685 and S686. This cleavage is performed by a host cell surface protease, furin, and is an important attribute in explaining the virulence of SARS-CoV-2 compared to other human 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.[42]

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.

[42] 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,[43] 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.

[43] 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?

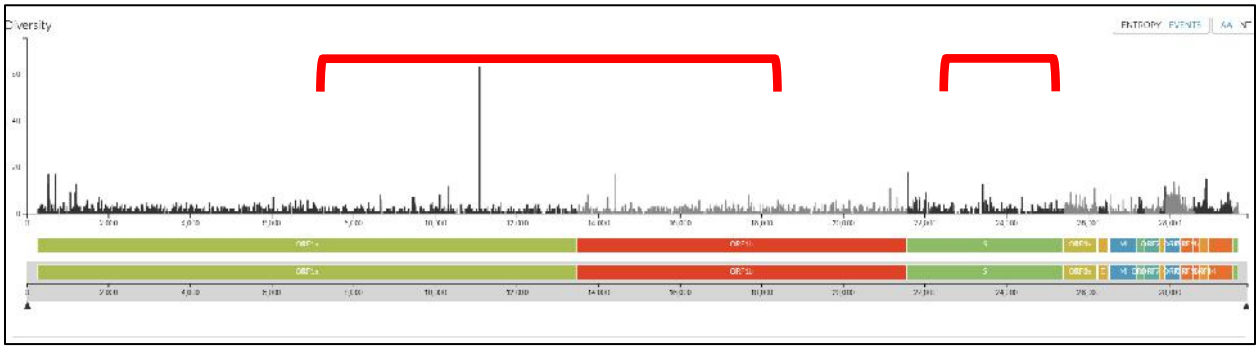
A simple explanation for these results would be an extreme criticality for the specific sequences of these regions with respect to infectivity. If a single amino acid change yielded a non-transmissible viral particle that strong negative purification process could explain the above results.

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."[44] 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.

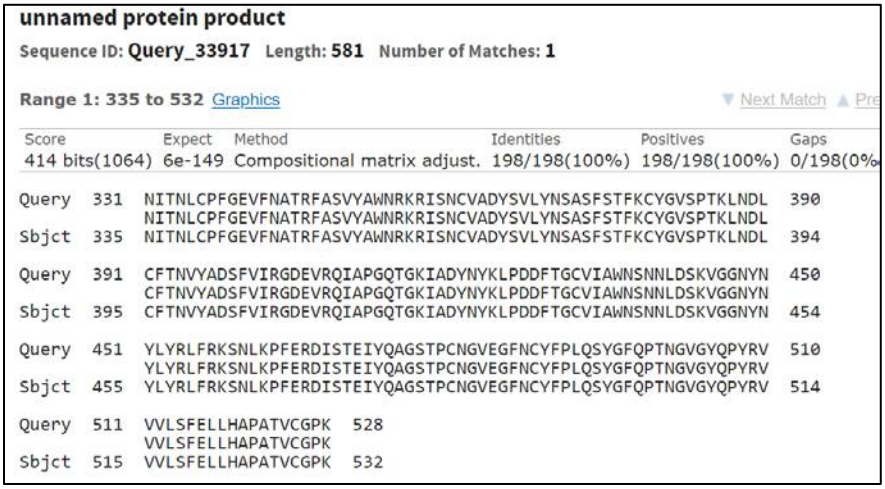
Codon modification, enhancement, or optimization is an example from synthetic biology in which the S/SN ratio is, by design, an anomaly when looked at through the lens of nature

[44] 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.

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[45] 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**

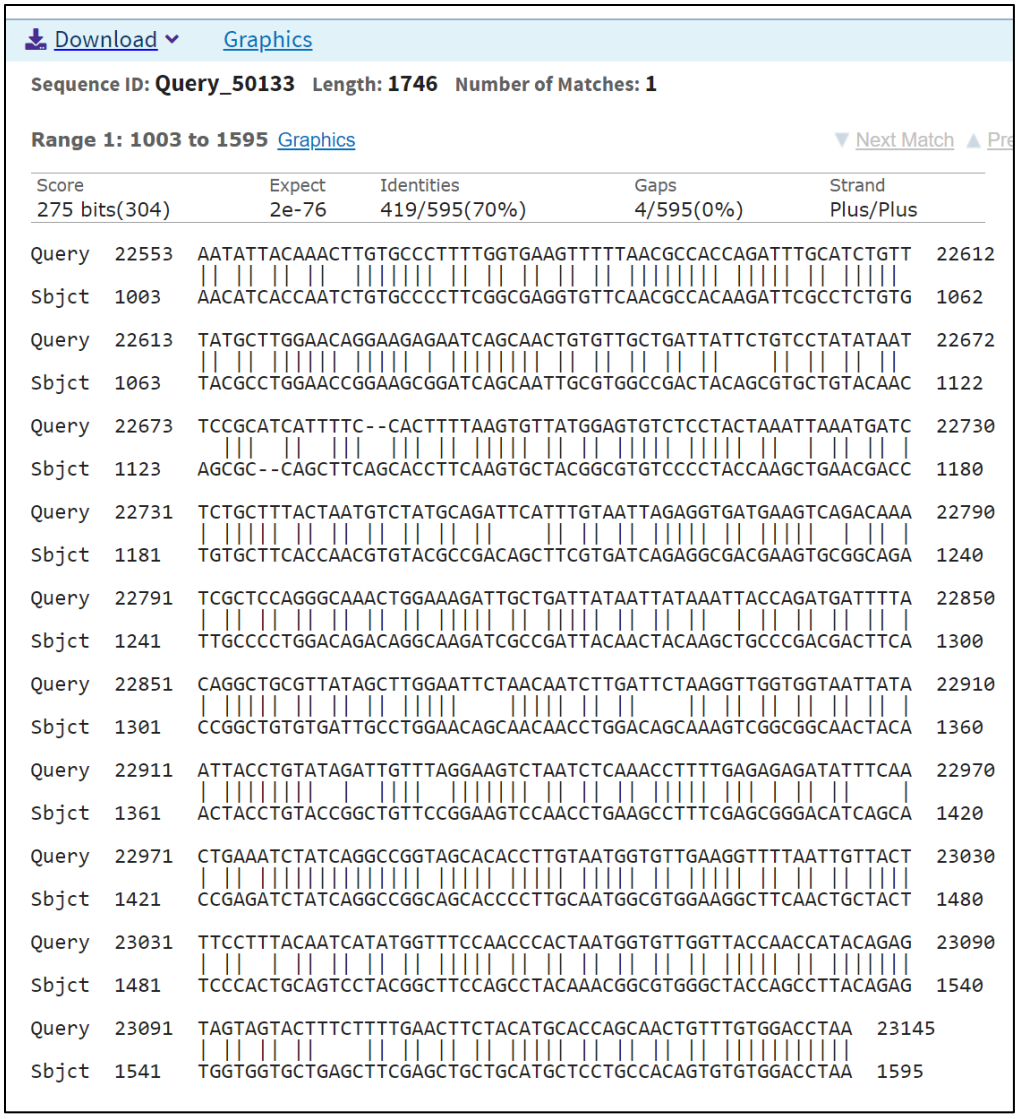
Range 1: 335 to 532 Graphics ▼ Next Match ▲ Pre

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:

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



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
```

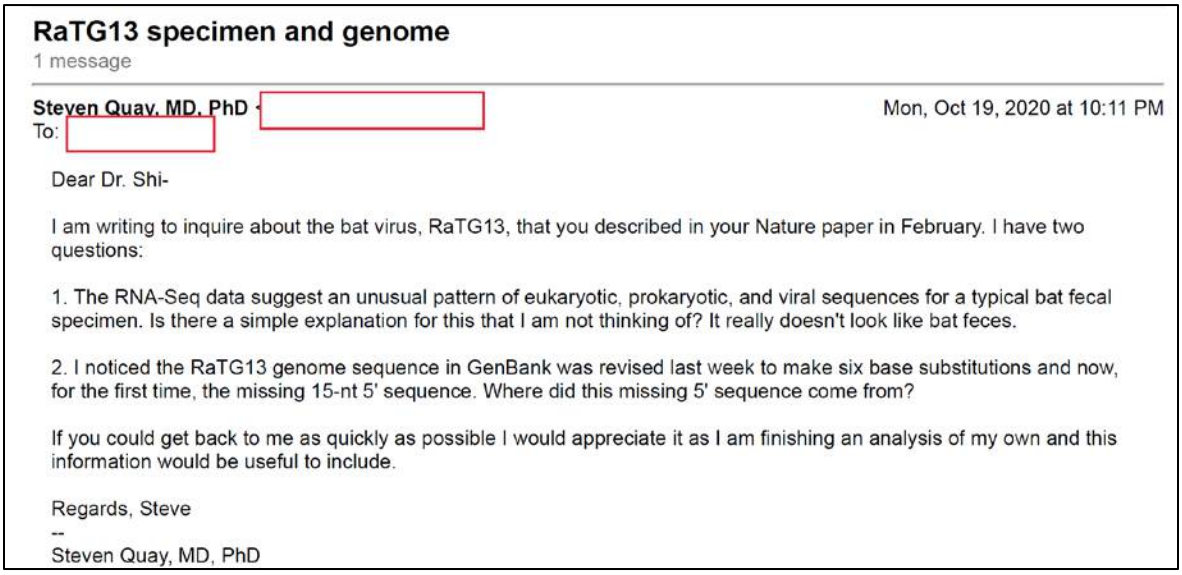
This is a situation in which there are 176 synonymous changes without a single non-synonymous change and is the genome signature of laboratory-derived synthetic biology. If these sequences were compared for phylogenetic divergence without the knowledge of their artificial construction, this synthetic 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.

October GenBank update. On October 13, 2020 the sequence for RaTG13 was updated. For the first time the first 15 nucleotides at the 5' end were present. However, these were not found in a blast of either the RNA-Seq raw reads or the Amplicons. The following email was sent to Dr. Shi asking for an explanation of the fecal specimen composition and the source for the 5' nt data.



RaTG13 specimen and genome
1 message

Steven Quay, MD, PhD [redacted] Mon, Oct 19, 2020 at 10:11 PM
To: [redacted]

Dear Dr. Shi-

I am writing to inquire about the bat virus, RaTG13, that you described in your Nature paper in February. I have two questions:

1. The RNA-Seq data suggest an unusual pattern of eukaryotic, prokaryotic, and viral sequences for a typical bat fecal specimen. Is there a simple explanation for this that I am not thinking of? It really doesn't look like bat feces.

2. I noticed the RaTG13 genome sequence in GenBank was revised last week to make six base substitutions and now, for the first time, the missing 15-nt 5' sequence. Where did this missing 5' sequence come from?

If you could get back to me as quickly as possible I would appreciate it as I am finishing an analysis of my own and this information would be useful to include.

Regards, Steve
--
Steven Quay, MD, PhD

At the time of this writing a response has not been received.

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.[46] Among primates, 18/19 have an ACE2 receptor which is

[46] 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

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.[47]

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 (*Eptesicus fuscus*),[48] Lander's horseshoe bat (*Rhinolophus landeri*), and Daubenton's bat (*Myotis daubentonii*) could not be infected with SARS-CoV-2.[49]

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).[50] 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.[51]

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.[52] However, the authors do not "(t)he

---

[47] 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

[48] J. Harcourt et al., Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. Emerg. Infect. Dis. 26, 1266–1273 (2020).

[49] 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).

[50] 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.

[51] 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).

[52] The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification

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 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.

Paper 2: The February 19, 2020 Lancet paper entitled: "Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19."

On February 19, 2020 *The Lancet* published a Correspondence entitled "Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19[53]" with 27 public health scientists from eight countries as authors. The statement seems to attempt to settle the question of the origin of SARS-CoV-2 and short circuit further debate, as the second sentence reads: "We stand together to strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin." It goes on to state: "Conspiracy theories do nothing but create fear, rumors, and prejudice that jeopardize our global collaboration in the fight against this virus."

The letter provided an open solicitation for support and at this time has been signed by at over 20,300 people, as if to purport that science can be advanced through polling and the democratic process.[54] While it is a truism that conspiracy theories have no place in the academia, legitimate debate should not be foreclosed.

The statement itself provides a more nuanced discussion of the evidence for a zoonotic origin and contains 14 references, eight of which contain data about the COVID-19 pandemic and six of which are governmental policy statements without new data, background articles from 2003 and 2004 on zoonotic diseases, or a virus naming statement by the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses, which is responsible for developing the official classification of viruses and taxa naming (taxonomy) of the Coronaviridae family. The eight articles with data were written at the end of January or early February, when there were fewer than 10,000 patients.

[53] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30418-9/fulltext#back-bib1

[54] This is reminiscent of the story attributed to Albert Einstein by Stephen Hawkins in his *Brief History of Time.* According to Hawkins, a book was published in 1930 in pre-war Germany entitled, "One Hundred Authors Against Einstein." When he was asked about the book Einstein is reported to have retorted, "If I were wrong, then one would have been enough!"

An analysis of the evidence for a zoonotic source given in support of the above Statement is contained in Text-Table here. The analysis shows there was very little actual data available at the time to permit reaching such a definitive conclusion. There was also the absence of data or discussion that could support a laboratory origin.

Reference	Statements concerning origin of SARS-CoV-2	Response to statements
1.Gorbalenya AE Baker SC Baric RS et al. Severe acute respiratory syndrome-related coronavirus: the species and its viruses—a statement of the Coronavirus Study Group. bioRxiv. 2020; (published online Feb 11. DOI: 2020.02.07.937862 (preprint).)	A naming statement about SARS-CoV-2. The emergence of SARS-CoV-2 as a human pathogen in December 2019 may thus be perceived as completely independent from the SARS-CoV outbreak in 2002–2003. With respect to novelty, SARS-CoV-2 differs from the two other zoonotic coronaviruses, SARS-CoV and MERS-CoV, introduced to humans earlier in the twenty-first century.	Does not provide data on a potential zoonotic source.
2.Zhou P Yang X-L Wang X-G et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; (published online Feb 3.)	The sequences of 2019-nCoV BetaCoV/Wuhan/WIV04/2019 among patient specimens are almost identical and share 79.6% sequence identity to SARS-CoV. Furthermore, we show that 2019-nCoV is 96% identical at the whole-genome level to a bat coronavirus. Pairwise protein sequence analysis of seven conserved non-structural proteins domains show that this virus belongs to the species of SARSr-CoV. The close phylogenetic relationship to RaTG13 provides evidence that 2019-nCoV may have originated in bats.	The bat genome identity of 96% described here, coupled with the known mutation rate of SARS-CoV-2 of about 26/year, implies a **lowest common ancestor about 44 years ago.**

3.Lu R Zhao X Li J et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; (published online Jan 30.)	Genome sequences of 2019-nCoV sampled from nine patients who were among the early cases of this severe infection are almost genetically identical, which suggests very recent emergence of this virus in humans and that the outbreak was detected relatively rapidly. 2019-nCoV is most closely related to other betacoronaviruses of bat origin, indicating that these animals are the likely reservoir hosts for this emerging viral pathogen.	Figure 1A shows 8 sequences and the concensus sequence. These 8 sequences show 3 with 0 mutations, 2 with 1 mutation, 3 with 2 mutations, and none with more than 2 mutations. Based on current estimates of 1 mutation per human passage, these are at most two human-to-human transfers apart. Importantly, there is no background diversity as would be seen in two or more resevoir-to-human events. Fig 2 states strain Bat-SL-CoVZC45 is 87.6% sequence identity to the human virus, which means a difference of about 3700 mutations or over 70 years from lowest common ancestor.
4.Zhu N Zhang D Wang W et al. A novel coronavirus from patients with pneumonia in China, 2019. NEJM. 2020; (published online Jan 24.)	"more than 85% identity with a bat SARS-like CoV (bat-SL-CoVZC45, MG772933.1) genome published previously. Since the sequence identity in conserved replicase domains (ORF 1ab) is less than 90% between 2019-nCoV and other members of betacoronavirus, the 2019-nCoV — the likely causative agent of the viral pneumonia in Wuhan — is a novel betacoronavirus belonging to the	A >85% identity with a bat coronavirus means the **human and bat virus have over 70 years to LCA.**

	sarbecovirus subgenus of Coronaviridae family."	
5.Ren L Wang Y-M Wu Z-Q et al. Identification of a novel coronavirus causing severe pneumonia in humans: a descriptive study. Chin Med J. 2020; (published online Feb 11.)	All five patients have sequence homology of 99.8% to 99.9%. These isolates showed 79.0% nucleotide identity with the sequence of SARS-CoV (GenBank NC_004718) and 51.8% identity with the sequence of MERS-CoV (GenBank NC_019843). The virus is closest to a bat SARS-like CoV (SL-ZC45, GenBank MG772933) with 87.7% identity, but is in a separate clade. **Surprisingly, RNA-dependent RNA polymerase (RdRp), which is the most highly conserved sequence among different CoVs, only showed 86.3% to 86.5% nt identities with bat SL-CoV ZC45.**	Similar to reference 3 comments. Lack of conserved sequencing of the most highly conserved sequence with bat coronavirus would **suggest a non-bat source.**
6.Paraskevis D Kostaki EG Magiorkinis G Panayiotakopoulos G Tsiodras S Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol. 2020; (published online Jan 29.)	A BLAST search of 2019-nCoV middle fragment revealed no considerable similarity with any of the previously characterized corona viruses. Bat_SARS-like coronavirus sequences cluster in different positions in the tree, suggesting that they are recombinants, and thus that the 2019-nCoV and RaTG13 are not recombinants. Codon usage analyses can resolve	**The middle segment with no similarity to other corona viruses is about 40% of the entire genome. I agree SARS-CoV-2 is not a recombinant of RaTG13. I agree, codon usage analysis here supports the furin binding site insertion as having been invented de novo. A recent recombination event is not necessary for a**

	the origin of proteins with deep ancestry and insufficient phylogenetic signal or **invented de novo**. Our study rejects the hypothesis of emergence as a result of a recent recombination event. Notably, the new coronavirus provides a new lineage for almost half of its genome, with **no close genetic relationships to other viruses** within the subgenus of sarbecovirus. This genomic part comprises half of the spike region encoding a multifunctional protein responsible also for virus entry into host cells	**laboratory derived theory of origin. Statements do not advance a zoonotic origin.**
7.Benvenuto D Giovanetti M Ciccozzi A Spoto S Angeletti S Ciccozzi M The 2019-new coronavirus epidemic: evidence for virus evolution. J Med Virol. 2020; (published online Jan 29.)	The epidemic originated in Wuhan, China. A phylogenetic tree has been built using the 15 available whole genome sequences of 2019-nCoV, 12 whole genome sequences of 2019-nCoV, and 12 highly similar whole genome sequences available in gene bank (five from the severe acute respiratory syndrome, two from Middle East respiratory syndrome, and five from bat SARS-like coronavirus). >97% maximum likelihood match to Bat SARS-like virus 2015 (Fig 1) is noted. The SARS and MERS viruses are excluded as a source of SARS-CoV-2. These results do not	A 3% genome **distance from the noted bat virus to human is about 34 years** at 26 mutations per year, the in-human mutation rate. Predicted a future mutation like the D614G mutation which is more infective.

	exclude the fact that **further mutation due to positive selective pressure, led by the epidemic evolution, could favor an enhancement of pathogenicity and transmission of this novel virus.**	
8.Wan Y Shang J Graham R Baric RS Li F Receptor recognition by novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS. J Virol. 2020; (published online Jan 29.)	Based on predicted RBD-host ACE2 receptor affinities, civet, mice, and rats are fuled out as source species. Pigs, ferrets, cats, and **nonhuman primates** contain largely favorable 2019-nCoV-contacting residues in their ACE2. SARS-CoV was isolated in wild palm civets near Wuhan in 2005, and its RBD had already been well adapted to civet ACE2.	The **potential nonhuman primate ACE2 usage is noted. Consistent with a laboratory origin from VERO cells**, a monkey kidney cell line. It expresses an ACE2 that permits SARS-CoV-2 infection, making it a possible source for the virus. A common tissue culture cell line forSARS virus research.
9.US Center for Disease Control and Prevention Coronavirus disease 2019 (COVID-19) situation summary. https://www.cdc.gov/coronavirus/2019-nCoV/summary.html Date: Feb 16, 2020 Date accessed: February 8, 2020	Rarely, animal coronaviruses can infect people and then spread between people such as with MERS-CoV, SARS-CoV, and now with this new virus, named SARS-CoV-2. The SARS-CoV-2 virus is a betacoronavirus, like MERS-CoV and SARS-CoV. All three of these viruses have their origins in bats. The sequences from U.S. patients are similar to the one that China initially posted, suggesting a likely single, recent emergence of this virus from an animal reservoir.	There are no data to support these statements about bats as the source for SARS-CoV-2.

10.Andersen KG Rambaut A Lipkin WI Holmes EC Garry RF The proximal origin of SARS-CoV-2. http://virological.org/t/the-proximal-origin-of-sars-cov-2/398 Date: Feb 16, 2020 Date accessed: February 17, 2020	See Table 2.	See Table 2.
11.Bengis R Leighton F Fischer J Artois M Morner T Tate C The role of wildlife in emerging and re-emerging zoonoses. Rev Sci Tech. 2004; 23: 497-512	In one pattern, actual transmission of the pathogen to humans is a rare event but, once it has occurred, human-to-human transmission maintains the infection for some period of time or permanently. Some examples of pathogens with this pattern of transmission are human immunodeficiency virus/acquired immune deficiency syndrome, influenza A, Ebola virus and severe acute respiratory syndrome.	This 2004 paper describes the pattern of rare animal-to-human transmission followed by human-to-human spread as an example of the SARS virus. It does not address the origin of SARS-CoV-2.
12.Woolhouse ME Gowtage-Sequeria S Host range and emerging and reemerging pathogens. Emerg Infect Dis. 2005; 11: 1842-1847	Emerging and reemerging pathogens are disproportionately viruses, with 37% being RNA viruses. Emerging and reemerging pathogens more often are those with broad host ranges that often encompass several mammalian orders and even nonmammals. For pathogens that are minimally transmissible within human populations (R0 close to 0), outbreak size is determined largely by the number of introductions from the reservoir. For pathogens that are highly transmissible within human populations	This 2005 article has good general information about looking broadly for the reservoir species(s), identifies RNA viruses as a major source of human epidemics, predicts a large outbreak size for a high Ro virus, but does address the origin of SARS-CoV-2 origin.

	(R0>>1), outbreak size is determined largely by the size of the susceptible population.	
13.NASEM The National Academies of Science Engineering and Medicine of the USA. NAS, NAE, and NAM presidents' letter to the White House Office of Science and Technology Policy. https://www.nationalacademies.org/includes/NASEM%20Response%20to%20OSTP%20re%20Coronavirus_February%206,%202020.pdf Date: Feb 6, 2020 Date accessed: February 7, 2020	The closest known relative of 2019-nCoV appears to be a coronavirus identified from bat-derived samples collected in China.4 The experts informed us that additional genomic sequence data from geographically- and temporally-diverse viral samples are needed to determine the origin and evolution of the virus. Samples collected as early as possible in the outbreak in Wuhan and samples from wildlife would be particularly valuable. Understanding the driving forces behind viral evolution would help facilitate the development of more effective strategies for managing the 2019-nCoV outbreak and for preventing future outbreaks.	Agree. If additional genomic sequence data is available from geographically- and temporally-diverse viral samples are needed to determine the origin and evolution of the virus this should be made publicly available.
14.WHO Director-General's remarks at the media briefing on 2019 novel coronavirus on 8 February 2020. https://www.who.int/dg/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020 Date: Feb 8, 2020 Date accessed: February 18, 2020	A general statement about the emerging pandemic without reference to the origin of SARS-CoV-2	There is no data about the origin of the pandemic.

In November 2020 the Watchdog group, US Right-to-Know, reported the following with respect to the *Lancet* article:[55]

"Emails obtained by U.S. Right to Know show that a statement in *The Lancet* authored by 27 prominent public health scientists condemning "conspiracy theories suggesting that COVID-19 does not have a natural origin" was organized by employees of EcoHealth Alliance, a non-profit group that has received millions of dollars of U.S. taxpayer funding to genetically manipulate coronaviruses with scientists at the Wuhan Institute of Virology."

"The emails obtained via public records requests show that EcoHealth Alliance President Peter Daszak drafted the Lancet statement, and that he intended it to "not be identifiable as coming from any one organization or person" but rather to be seen as "simply a letter from leading scientists". Daszak wrote that he wanted "to avoid the appearance of a political statement."

A separate, worrisome article entitled, "Peter Daszak's EcoHealth Alliance Has Hidden Almost $40 Million In Pentagon Funding And Militarized Pandemic Science,[56]" seems to indicate a serious conflict of interest with respect to Dr. Daszak's participation in any investigations on the origin of SARS-CoV-2.

Paper 3: The March 17, 2020 article in *Nature Medicine* entitled "The proximal origin of SARS-CoV-2" by Andersen et al.[57, 58]

According to the journal, this article is in the 99th percentile (ranked 2nd) of the 312,683 tracked articles of a similar age in all journals and the 99th percentile (ranked 1st) of the 147 tracked articles of a similar age in *Nature Medicine.* The metrics also indicate it has been accessed over five million times. It is clearly the most cited paper and since its title and topic are the origin of the pandemic it clearly has an outsized influence on the topic.

The following statements form the evidence in the article of the natural origin of CoV-2:

- "While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses **predict that the interaction is not ideal** and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding. Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is **most likely the result of natural selection on a human or human-like ACE2** that permits another optimal binding solution to arise. **This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation."** [emphasis added.]

[55] https://usrtk.org/biohazards-blog/ecohealth-alliance-orchestrated-key-scientists-statement-on-natural-origin-of-sars-cov-2/
[56] https://www.independentsciencenews.org/news/peter-daszaks-ecohealth-alliance-has-hidden-almost-40-million-in-pentagon-funding/
[57] https://www.nature.com/articles/s41591-020-0820-9
[58] Two non-peer reviewed analyses are included here because they provide a nearly line-by-line analysis. They unfortunately include occasional colorful language but the content is worth noting: https://harvardtothebighouse.com/2020/03/19/china-owns-nature-magazines-ass-debunking-the-proximal-origin-of-sars-cov-2-claiming-covid-19-wasnt-from-a-lab/ ; https://www.youtube.com/watch?v=HmSCMb8Nds4

  - A later analysis of over 3800 possible substitutions of amino acids in a 200 amino acid receptor binding region, much larger than the small, selective region referred to in this paper, shows that CoV-2 is 99.5% optimized for binding to the ACE-2 receptor. This near perfect binding has never been seen before in a recent interspecies transmission jump.

- "Polybasic cleavage sites have not been observed in related 'lineage B' betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have those sites and predicted O-linked glycans. Given the level of genetic variation in the spike, **it is likely that SARS-CoV-2-like viruses with partial or full polybasic cleavage sites will be discovered in other species.**" [emphasis added.]

  - As of the writing of this manuscript no other lineage B (sarbecovirus) has been found to have a furin site. In addition, the furin site of CoV-2 has the unusual -CGG-CGG- codon dimer, which has never been seen in an analysis of 58 other sarbecoviruses, that is, 580,000 codons. Since recombination between subgenera of beta coronaviruses is rare, or unknown, there is no source for the CGG-CGG dimer via a natural recombination event.

- "The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals."

  - It is curious why the above statement did not lead to a hypothesis somewhere in the article about a similar mechanism on CoV-2, a clear indication of a laboratory origin.

- "It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-CoV-like coronavirus."

  - This conclusory statement is unsupported my evidence.

- "Furthermore, if genetic manipulation had been performed, one of the several reverse-genetic systems available for betacoronaviruses would **probably have been used**. However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone." [emphasis added.]

  - There is no explanation for why a prior backbone would necessarily be used. All synthetic biology chimera coronaviruses created in the past as published in prior papers have each used a unique backbone with no particular pattern in backbone selection. Each backbone was selected for the particular needs of those current experiments. This non-repeating prior pattern of reverse-genetic systems makes the above statement untenable. And with 16,000+ reported coronavirus specimens at the WIV it entirely reasonable a non-published virus could have been used.

- "Natural selection in an animal host before zoonotic transfer. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for binding to human ACE2, **an animal host would probably have to have a high population density (to allow natural selection to proceed efficiently**) and an ACE2-encoding gene that is similar to the human ortholog." [emphasis added.]

    - The paragraph discusses the pangolin as the possible intermediate host but at the time of this manuscript the coronavirus data from pangolins has been discredited. This author agrees with statement that selection of the two unique features of CoV-2 require a high population density of the animal host. Of course, in the laboratory the animal hosts for either *in vitro* cell culture experiments or in animal experiments are a single species at high density.

- Natural selection in humans following zoonotic transfer. "It is possible that a progenitor of SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during **undetected human-to-human transmission**. Once acquired, these adaptations would enable the pandemic to take off and produce a sufficiently large cluster of cases to trigger the surveillance system that detected it." [emphasis added.]

- "Studies of banked human samples could provide information on whether such cryptic spread has occurred. Further serological studies should be conducted to determine the extent of prior human exposure to SARS-CoV-2."

    - As will be shown in later sections, this prior undetected human-to-human transmission would be evident in archived specimens from before the fall of 2019. In both SARS-CoV-1 and MERS, this prior seroconversion averaged about 0.6% with almost 5% among workers exposed to the intermediate hosts. At the time of the writing of this manuscript, in limited sampling of archived specimens there has been no seroconversion detected. The author believes there are thousands of archived specimens from Wuhan taken in the fall of 2019 and these should be immediately examined for evidence of seroconversion. Since finding seroconversion among these specimens would be strong evidence for a zoonotic origin and not a laboratory accident, the absence of any information from China on this important evidence is hard to understand.

- Selection during passage. "Basic research involving passage of bat SARS-CoV-like coronaviruses in cell culture and/or animal models has been ongoing for many years in biosafety level 2 laboratories across the world, and there are documented instances of laboratory escapes of SARS-CoV. We must therefore examine the possibility of an inadvertent laboratory release of SARS-CoV-2."

- "In theory, it is possible that SARS-CoV-2 acquired RBD mutations during adaptation to passage in cell culture, as has been observed in studies of SARS-CoV."

- "New polybasic cleavage sites have been observed only after prolonged passage of low-pathogenicity avian influenza virus in vitro or in vivo. Furthermore, a hypothetical generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with very high genetic similarity, **which has not been described.** Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to those of humans, **but such work has also not previously been described**." [emphasis added.]

    - The authors correctly describe a method for CoV-2 to have been generated in the laboratory and then dismiss it because the work has not been published previously. As active scientists themselves, the authors must know how disingenuous this sounds. Almost by definition elite scientists, like Dr. Shi of the WIV, work in secret until the publication of any given line of research. As the say, the absence of evidence cannot be used as evidence of its absence.

    - A peer-reviewed paper[59] entitled, "Might SARS-CoV-2 Have Arisen via Serial Passage through an Animal Host or Cell Culture? A potential explanation for much of the novel coronavirus' distinctive genome," provides a compelling argument that serial passage in the laboratory might indeed have been the manner in which CoV-2 acquired many of its devastating traits.

- "Although the **evidence shows that SARS-CoV-2 is not a purposefully manipulated virus**, it is currently impossible to prove or disprove the other theories of its origin described here. However, **since we observed all notable SARS-CoV-2 features, including the optimized RBD and polybasic cleavage site, in related coronaviruses in nature, we do not believe that any type of laboratory-based scenario is plausible**." [emphasis added.]

    - This author could identify no prior evidence in the paper to warrant saying it is not a purposefully manipulated virus. There is also no evidence that would point to a purposely manipulated virus.

    - The evidence in the paper shows that no prior zoonotic interspecies transmission has ever had an RBD as optimized as the CoV-2 RBD for the human ACE2. The evidence also shows that there is no natural source for the polybasic cleavage site (PCS). No other member of the subgenera to which CoV-2 belongs has a PCS. Since these are the only coronaviruses from which recombination could supply a polybasic cleavage site, the data in this paper refutes the natural origin.

[59] https://onlinelibrary.wiley.com/doi/full/10.1002/bies.202000091

- The belief statement concerning a laboratory-based scenario would be closer to the evidence if it was professed with, “despite evidence which is consistent with a laboratory-based scenario.”

Based on the author’s analysis of the paper, the following email was sent to the lead author:

Gmail Steven Quay, MD, PhD

SARS-CoV-2 origin

Steven Quay, MD, PhD Mon, May 25, 2020 at 7:14 PM
To:

Dr. Andersen-

I read with interest your paper titled, The proximal origin of SARS-CoV-2, in which you conclude this is a natural, zoonotic-sourced infection. Three different approaches of analysis that I have done do not support this conclusion. Can you comment please:

1. The furin cleavage site insert has the unusual codon usage for the RR dimer of CGG-CGG. As you probably know, the frequency of this codon usage in the SARS-CoV-2 genome is 0.09. So having two next to each other is not likely as a random event. As support for the unlikeliness of these codons, using GISAID data, by March there is evidence of the third G being mutated out for either A or T at three-times the rate of the background mutation rate, 26/year from Nextstrain.org. Since codon usage in coronaviruses are not greatly influenced by the host it resides in, this means a jump to humans must have been in a host which did not have even a few months history with the virus, otherwise the terminal G would have been purified out. On the other hand, most laboratory use optimized codon primers and kits use CGG routinely; including in primers in published papers from the Wuhan Institute of Virology. So a laboratory source for a gain-of-function furin cleavage site would probably use these codons.

2. In the over 16,000 genomes in GISAID there is not one example of posterior diversity. With MERS, 93% of sequenced genomes did not pass through the index case but represented separate reservoir host to human jumps and this was apparent within 60 days of the index case. They collectively showed the most recent common ancestors among themselves was over 12 months before the index case. With SARS-CoV-2 it is acting like a 'pure culture' growth from the index case outward with no evidence of a reservoir host in the background. This would be the case for a laboratory acquired infection.

3. If you use a map of Wuhan and overlay the first four hospitals that saw cases with a map of the Metro system, you see that the hospitals straddle Line 2, which runs approximately east to west, carries 1,000,000 people a day, and is the Metro line with stops closest to both the Wuhan Institute of Virology and the original wet market that was considered an early source for the infection. There are 11 Metro lines in Wuhan, hundreds of stops on those lines, and over a dozen hospitals spread out over the city. I am working with a UCLA statistician to perform tests about the probability of this being simply an accident of statistics but the gestalt is, it does not look like a chance occurrence. But it is consistent with someone getting infected in the lab, riding Line 2 for a few days, and off you go.

It might be a truism to say that the six proven cases of laboratory derived SARS escapes occurred in big cities, Beijing, Singapore, or Taipei where the labs are located. But if you follow it with the fact that MERS and SARS, both proven as true zoonotic sources, on other hand began in rural settings in China and the Middle East, respectively. I am not sure why this obvious correlation was not at least pointed out in your paper and then addressed with a cogent argument.

I look forward to hearing your thoughts.

Regards, Steve

Soon after this email was written Dr. Andersen blocked the author from following his Twitter account. A reply to the above email was never received.

Conclusion. Three high visibility papers were published between January and May 202 which purported to settle the question of the origin of SARS-CoV-2 as a zoonotic transmission and not a laboratory accident. The analysis above concludes that these papers are not persuasive. The

author has elected to not use evidence within these papers to change the prior likelihood of a zoonotic versus laboratory origin. They are presented here as neutral evidence that supports neither theory.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

Evidence. SARS-like infections among employees of the Wuhan Institute of Virology in the fall of 2019

The State Department of the United States issued the following statement on January 15, 2021[60]:

"1. Illnesses inside the Wuhan Institute of Virology (WIV):

- The U.S. government has reason to believe that several researchers inside the WIV became sick in autumn 2019, before the first identified case of the outbreak, with symptoms consistent with both COVID-19 and common seasonal illnesses. This raises questions about the credibility of WIV senior researcher Shi Zhengli's public claim that there was "zero infection" among the WIV's staff and students of SARS-CoV-2 or SARS-related viruses."

There is no additional evidence to support either parties position in the above statement. The U.S. Government statement would be considered hearsay in a court of law and probably not admissible. The veracity of Dr. Shi's statement above could be called into question due to other inconsistencies in some of her testimony, as reported elsewhere in this document.

At this time, the above evidence cannot be used to change the likelihood of either theory about the origin of SARS-CoV-2. The statement is kept within this analysis with the hope that in the future new information will come to light that could make this evidence a useful addition to the overall analysis.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

[60] https://2017-2021.state.gov/fact-sheet-activity-at-the-wuhan-institute-of-virology//index.html

Evidence. A Bayesian Analysis of one aspect of the SARS-CoV-2 origin, where the first recorded outbreak occurred, increases the probability of a laboratory origin.

Introduction. The two competing hypotheses of the origin of SARS-CoV-2 as a natural, zoonotic spillover event versus a laboratory-acquired infection (LAI) or other laboratory accident each had supporting evidence from the very beginning of the pandemic.

On the one hand, about 40% of early patients with COVID-19 had an association with the Hunan Seafood Market in Wuhan. Since this mirrored SARS-CoV-1, where markets selling civet cats were determined to be the origin of that human epidemic, the natural origin hypothesis seemed logical. The Chinese CDC have now ruled out the market as a source for the outbreak.

On the other hand, the laboratory origin hypothesis also had an early beginning with the fact that the outbreak began adjacent to the only high security, BSL-4 laboratory in all of China, and one of the top coronavirus research centers in the world, was the Wuhan Institute of Virology (WIV). The hospitals of the first COVID patients were very close to the WIV.

This evidence statement is taken from an article applying a Bayesian analysis to the hypothesis that the proximal origin of SARS-CoV-2 was an uncontrolled[61] release from a laboratory using, as evidence, one aspect of the SARS-CoV-2 origin story — where the first recorded outbreak occurred.[62]

Hypothesis: The first recorded outbreak of SARS-CoV-2 in the human population occurred in a city that is also home to a virology laboratory that actively performs research on closely related viruses.

In this case, the city is Wuhan, and the virology laboratory is run by the Wuhan Institute of Virology.

Analysis. This analysis set the likelihood of a laboratory escape (the prior probability the hypothesis was true) at three values, 0.01%, 0.1%, and 1.0%. The second term was the conditional probability of the evidence, given that the hypothesis is actually false. This was set at 0.01. Finally, the third term was the conditional probability of the evidence, given the hypothesis is true. This was set, biasing to the natural origin, at 0.71.

Results. The paper provides the three-by-three cube of results for the three parameters of interest.

The ardent sceptic's probability begins at 0.01% and the revised estimate is no more than 0.05% or 5/10000. It applies to someone who was initially very skeptical about a lab origin (0.01% probability), who believes there is no more than 51% chance that an uncontrolled release of a highly contagious disease would lead to a local outbreak, and who thinks there was at least a

[61] By using the term uncontrolled release, the author was specifically excluding from consideration the possibility that the pathogen was deliberately released from the laboratory.
[62] https://jonseymour.medium.com/a-bayesian-analysis-of-one-aspect-of-the-sars-cov-2-origin-story-where-the-first-recorded-1fbdcbea0a2b

10% chance that a natural outbreak of a virus native to Yunnan would have occurred in Wuhan before any place else.

On the other extreme, is the ardent believer who started with at least a 1% belief in a laboratory outbreak, is 100% certain that an uncontrolled laboratory release would result in a local outbreak and believes that the probability that a natural outbreak of a virus native to Yunnan would occur in Wuhan before any place else is less than 0.1%. The ardent believer's revised belief is that the probability that the Wuhan outbreak was caused by an uncontrolled laboratory release changes from 1% to at least 91%.

In the center, is the so-called "central" observer who accepts that the central values for each of the parameter ranges are reasonable estimates of the true values of the probability being estimated. The central observer started with an initially skeptical belief in the hypothesis of 0.1%, believes that average citizen in Wuhan was a likely as any other citizen of China to be the initial vector of the virus into the human population and believes that there is no more or less than a 71% chance that an uncontrolled release from a laboratory of a highly contagious pathogen such as SARS-CoV-2 would result in a local outbreak as opposed to an outbreak in some other location. The central observer's revised belief in the hypothesis is 6.8%. If the central observer began with a 1% belief in a laboratory origin, this analysis would change that to 41.8%.

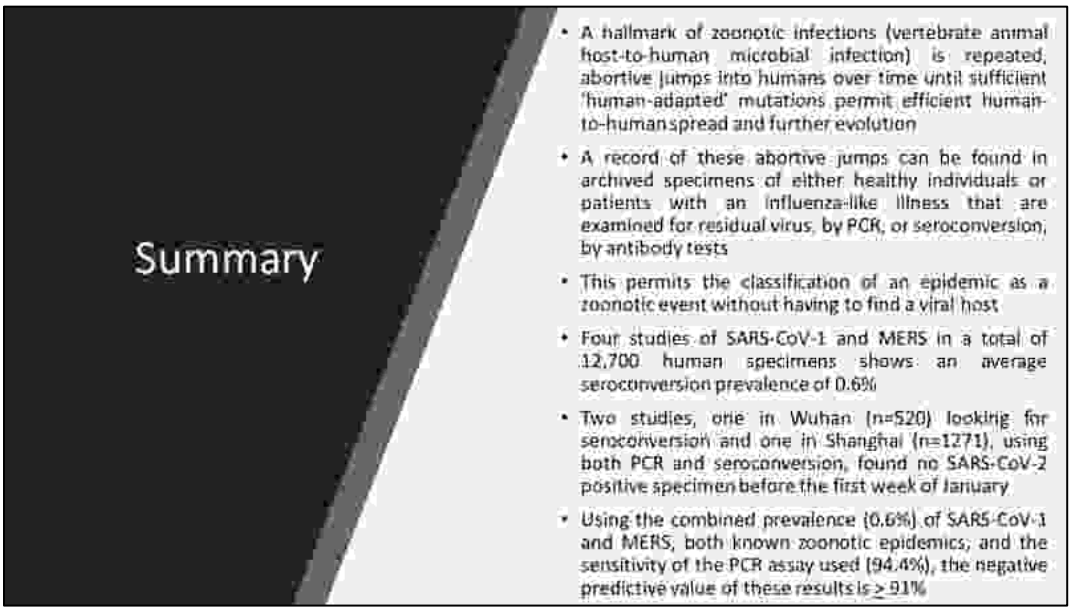
Conclusion. For purposes of this analysis and to be as conservative as possible, the assumptions will be that there is at least a 1% prior belief in a laboratory outbreak (because that was our starting probabilities), but there is no more than a 51% chance that an uncontrolled release of a highly contagious disease would lead to a local outbreak, and that there was at least a 10% chance that a natural outbreak of a virus native to Yunnan would have occurred in Wuhan before any place else. Using these assumptions, the initial likelihood of a 1% laboratory origin changes to 4.9%.

Starting likelihood from initial state: Zoonotic origin (98.8%) and laboratory origin (1.2%)

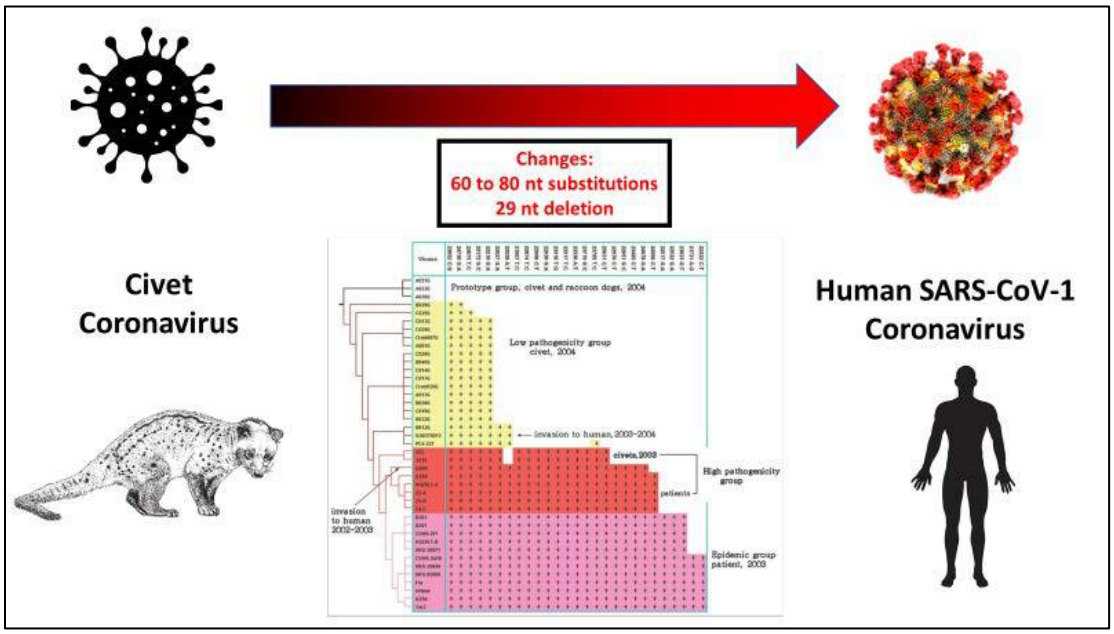
Adjusted likelihood: Zoonotic origin (95.1%) and laboratory origin (4.9%)

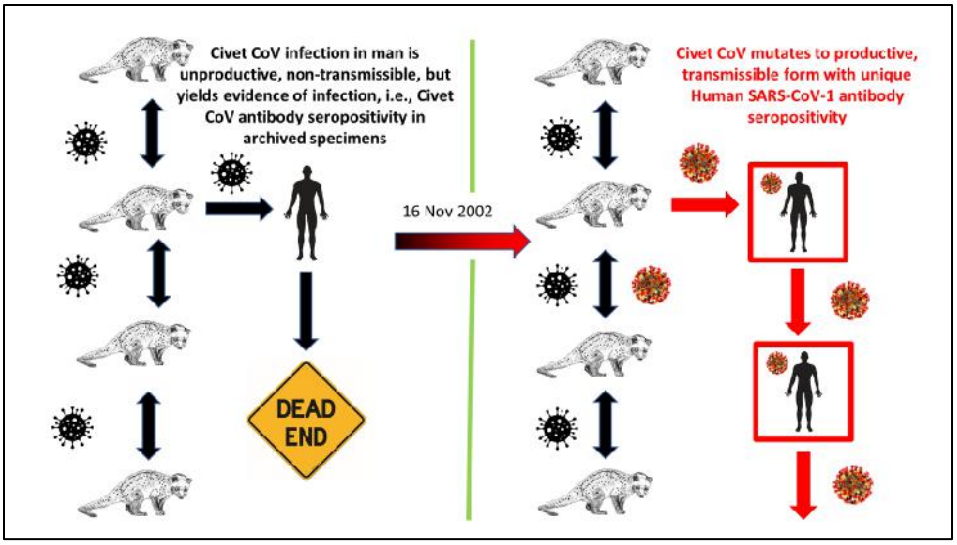
Evidence: Lack of seroconversion in Wuhan and Shanghai. Summary of evidence:

- A hallmark of zoonotic infections (vertebrate animal host-to-human microbial infection) is repeated, abortive jumps into humans over time until sufficient 'human-adapted' mutations permit efficient human-to-human spread and further evolution

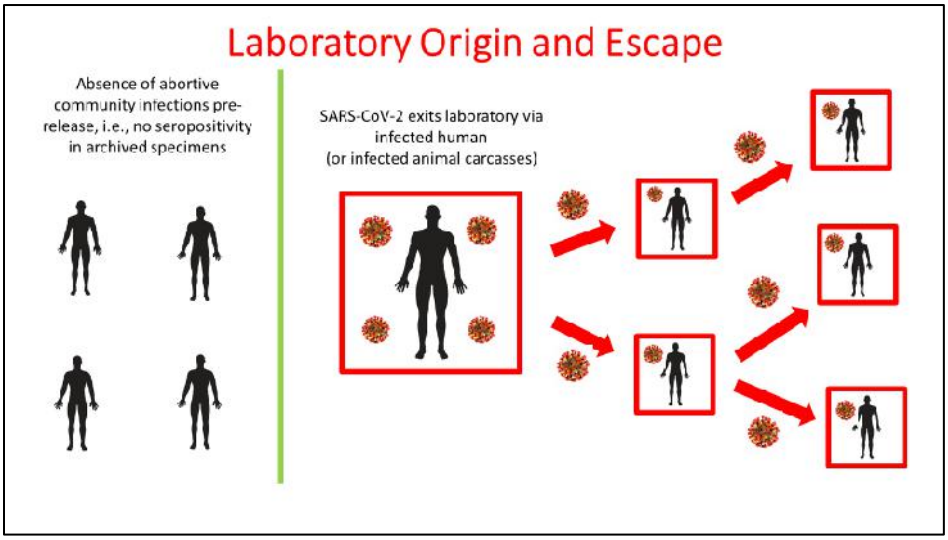


- A record of these abortive jumps can be found in archived specimens of either healthy individuals or patients with an influenza-like illness that are examined for residual virus, by PCR, or seroconversion, by antibody tests





- This permits the classification of an epidemic as a zoonotic event without having to find a viral host
- A laboratory accident is a situation in which there are no prior exposures within the human population as shown in the Figure below:



- Four studies of SARS-CoV-1 and MERS in a total of 12,700 human specimens shows an average seroconversion prevalence of 0.6%

SARS-related Virus Predating SARS Outbreak, Hong Kong

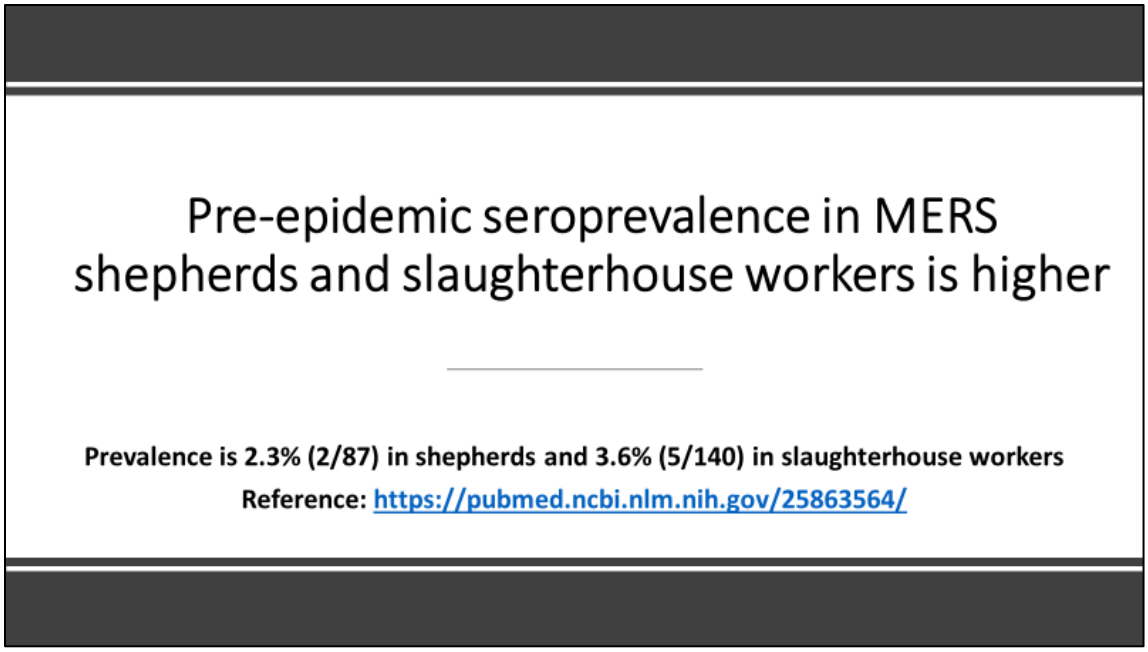
SARS-CoV-1 began in fall of 2002 in southern China

Patient Population	Serum samples collected in May 2001 from 938 healthy adults in Hong Kong	48 confirmed SARS patients diagnosed in February and March 2003 in Guangdong
Civet CoV > SARS-CoV-1 Seropositivity	13	0
SARS-CoV-1 > Civet CoV Seropositivity	4	48
Total	17 out of 938 = 1.8%	48 out of 48 = 100%

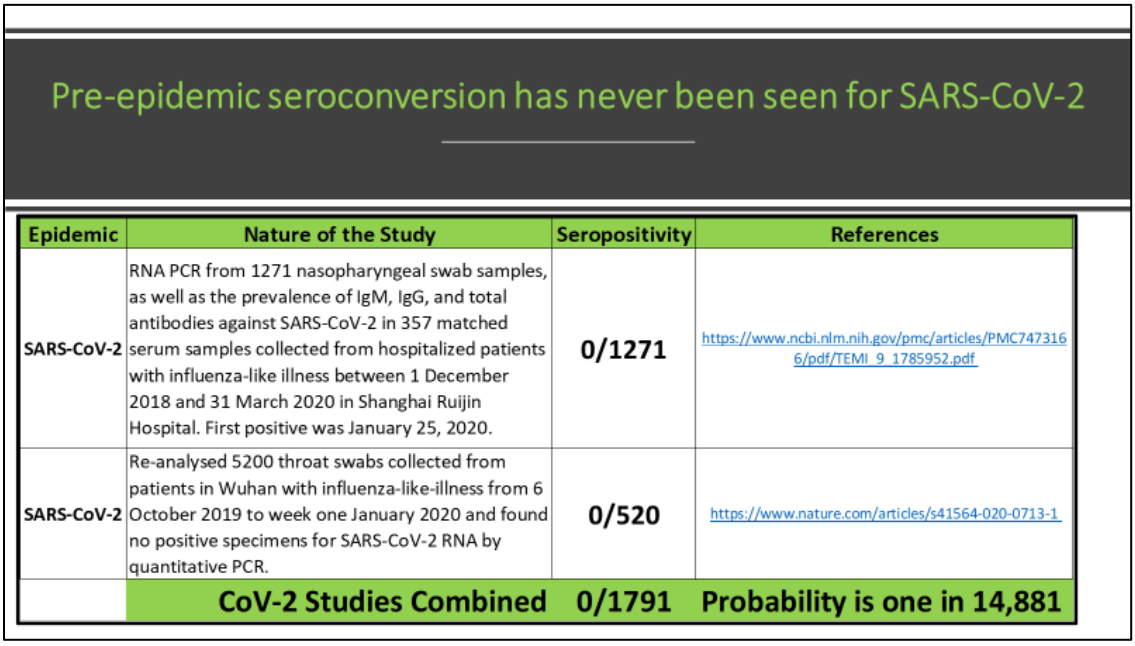
Pre-epidemic seroprevalence in the adult community

Prevalence is 0.6% for SARS-CoV-1 and MERS in 12,700 specimens

Epidemic	Nature of the Study	Seropositivity	Reference
SARS-CoV-1	Archived specimens from healthy adults in Hong Kong collected two years before CoV-1 were tested for Ab to civet or human CoV	17/938	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322899/
MERS	Archived human sera collected in 2011 was tested for MERS-CoV S1-specific antibodies by ELISA	1/90	https://www.sciencedirect.com/science/article/pii/S1876034120300010#fig0010
SARS-CoV-1	Serum specimens collected from military recruits from the People's Republic of China in 2002 were tested for SARS-CoV-1 antibodies.	11/1621	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1074388/
MERS	Between Dec 1, 2012, and Dec 1, 2013, 10,009 individual serum samples were tested for anti-MERS-CoV antibodies in regions without cases.	15/10,009	https://pubmed.ncbi.nlm.nih.gov/25863564/
SARS-CoV-1	Serum samples that were collected from 42 individuals during 2001-2002, before the SARS outbreak, and tested for IgG antibody against SARS-CoV.	28/42	https://arxiv.org/ftp/arxiv/papers/1305/1305.2659.pdf

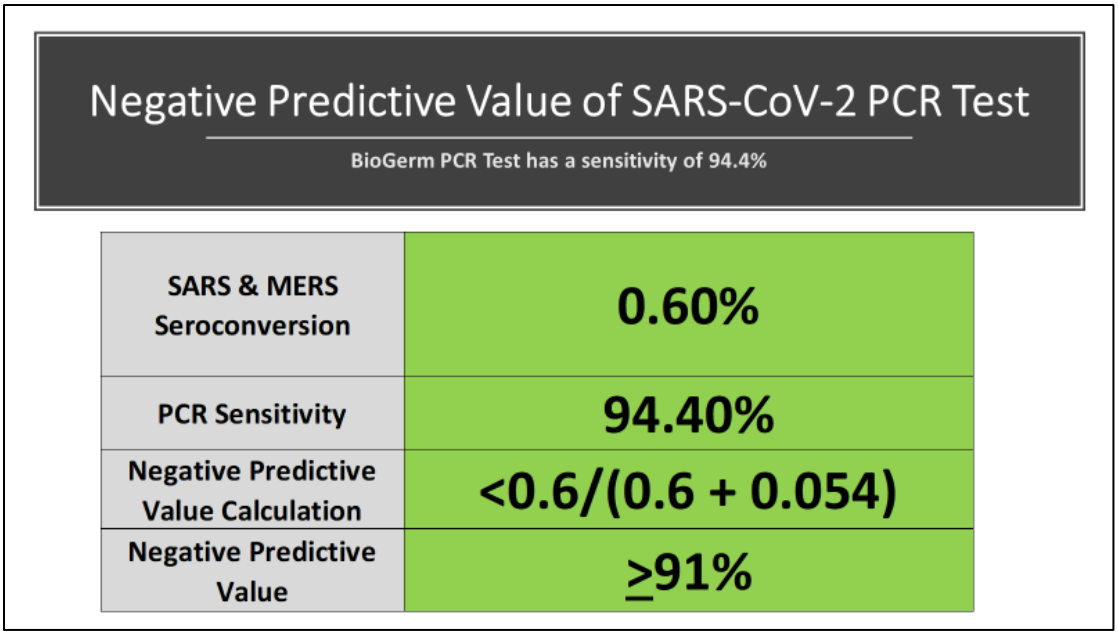


- Two studies, one in Wuhan (n=520) looking for seroconversion and one in Shanghai (n=1271), using both PCR and seroconversion, found no SARS-CoV-2 positive specimen before the first week of January



Epidemic	Nature of the Study	Seropositivity	References
SARS-CoV-2	RNA PCR from 1271 nasopharyngeal swab samples, as well as the prevalence of IgM, IgG, and total antibodies against SARS-CoV-2 in 357 matched serum samples collected from hospitalized patients with influenza-like illness between 1 December 2018 and 31 March 2020 in Shanghai Ruijin Hospital. First positive was January 25, 2020.	0/1271	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7473166/pdf/TEMI_9_1785952.pdf
SARS-CoV-2	Re-analysed 5200 throat swabs collected from patients in Wuhan with influenza-like-illness from 6 October 2019 to week one January 2020 and found no positive specimens for SARS-CoV-2 RNA by quantitative PCR.	0/520	https://www.nature.com/articles/s41564-020-0713-1
	CoV-2 Studies Combined	**0/1791**	**Probability is one in 14,881**

- Using the combined prevalence (0.6%) of SARS-CoV-1 and MERS, both known zoonotic epidemics, and the sensitivity of the PCR assay used (94.4%), the negative predictive value of these results is $\geq$ 91%



Here, the negative predictive value (NPV) represents the probability that a CoV-2 is not a zoonosis, given the negative seroconversion findings.

Subjective Discount Factor: 90% (a one in 10 chance this is wrong). This is a subjective value.

The change in origin likelihoods from this evidence and the calculations are shown in the Text-Table below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.951	0.049
Negative predictive value of lack of seroconversion	0.91	
Reduced by 90% Subjective Discount Factor	0.91 x 0.9 = 0.82	
Impact of this evidence	Reduces the likelihood of ZO by 82/18 or 4.6-fold. For every 100 tests, a true ZO would be seen 18 times and a non-ZO would be seen 82 times	
Impact of evidence calculation	0.951/4.6 = 0.207	
Normalize this step of analysis	0.207/(0.207 + 0.049) = 0.809	0.049/(0.207 + 0.049) = 0.191

Adjusted likelihood: Zoonotic origin (80.9%) and laboratory origin (19.1%)

Evidence: Lack of posterior diversity for SARS-CoV-2 compared to MERS and SARS-CoV-1

- The earliest stages of human CoV-1 and MERS infections were characterized by viral genome base diversity as expected for multiple, independent jumps from a large and diverse intermediate host population into humans.
- Combining MERS and CoV-1 studies, out of the earliest 255 human infections in which virus genome sequences are available, 137 could not be rooted in a prior human-to-human infection and so are attributed to an independent intermediate host-to-human infection.[63]
- That is about 54% non-human-to-human transmission.
- On the other hand, Ralph Baric has written[64] that CoV-2 is different: "SARS-CoV-2 probably emerged from bats, and early strains identified in Wuhan, China, showed limited genetic diversity, which suggests that the virus **may have been introduced from a single source**." [emphasis added.]
- With CoV-2, there are 249 viral genomes in GISAID from Hubei province, where Wuhan is located, collected between Dec 24, 2019 and Mar 29, 2020.
- From Dec 24, 2019 to November 2020, there are 1001 genomes sequenced from all of China and 198,862 worldwide.
- For CoV-2, every single genome sequence is rooted in the first sequence from the PLA Hospital in Wuhan.
- Not one case of posterior diversity.
- Using the frequency of non-rooted genome diversity seen with MERS and CoV-1, about 50:50 or a coin toss, the probability that CoV-2 is a zoonotic pandemic with 0/249 genomes is the chance of tossing a coin 249 times and getting heads every time!
- Mathematically that is nonexistent; specifically, one in 10 with 84 zeros.
- Since Wuhan had approximately 500,000 cases during the time interval of this sampling, the potential sampling error of testing only 249/500,000 or 0.05% is significant. This sampling error, while large, is unable to obliterate the overwhelming odds that this did not arise from an intermediate host in Wuhan.
- Therefore, to permit continued evidence analysis, this finding will be set at the boundary of customary statistical significance, a p-value of 0.05 or a 1 in 20 likelihood that this is zoonotic.

[63] https://elifesciences.org/articles/31257#abstract ; https://www.researchgate.net/publication/225726653_Molecular_phylogeny_of_coronaviruses_including_human_SARS-CoV ; https://science.sciencemag.org/content/300/5624/1394/tab-pdf ; https://pubmed.ncbi.nlm.nih.gov/14585636/ ; https://www.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.016378-0?crawler=true ; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7118731/

[64] https://www.nejm.org/doi/10.1056/NEJMcibr2032888

Detailed explanation

A fundamental difference between a laboratory and a non-laboratory acquired zoonotic disease, the imprint of phylogenetic diversity through pre-human spread within the source population, can be examined by the posterior diversity of human cases with no *a priori* knowledge of an intermediate host.

MERS. The MERS epidemic has been documented to have arisen from the initial jump from bats to camels, a three-to-five-year expansion within the camel population in which mutational diversity arose by random mistakes, and then a jump into humans. This model of spread predicts that there would, at some point, be additional jumps from other camels into other patients, and a pattern of "posterior diversity," would be found in the human specimens. If the COVID-19 pandemic arose by a similar mechanism the same pattern would be seen. The following Text-Table contains such data.

Phylogenetic Feature	MERS	SARS-CoV-2
Posteriority Diversity	28/30 (93%)	0
No Posteriority Diversity	2/30 (7%)	7666
Time from first patient to first example of posterior diversity	About 60 days	None at >120 days
Depth of posterior diversity to first patient	>365 days	None

The study of MERS noted above was published in 2013 in Lancet[65] in an article entitled, "Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study." Thirty specimens were used in the analysis. The features of a camel-to-human zoonotic epidemic are easily identified. Specimens taken within sixty days of the first patient, "Patient Zero," began to show a background diversity that could not be traced back through Patient Zero. The analysis of all thirty, in fact, documented that 93% were transmitted directly from the camel intermediate reservoir. And looking only at the "background" diversity permitted a calculation of the last common ancestor for the spread within the camel population of over 365 days.

A study of SARS-CoV-2[66] available May 5, 2020 and entitled, "Emergence of genomic diversity and recurrent mutations in SARS-CoV-2," looked at 7666 patient specimens from around the world for phylogenetic diversity. The authors state: "There is a robust temporal signal in the data, captured by a statistically significant correlation between sampling dates and 'root-to-tip' distances for the 7666 SARS-CoV-2 ($R^2 = 0.\ 20$, $p < .001$). Such positive association between sampling time and evolution is expected to arise in the presence of measurable evolution over the timeframe over which the genetic data was collected." This conclusion also argues against a

[65] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898949/
[66] https://www.sciencedirect.com/science/article/pii/S1567134820301829

MERS-like pattern of posterior diversity. In fact, the 95% upper bound for the probability of no posterior diversity being seen in SARS-CoV-2, given the data in MERS, is 3.9×10^{-4}.

The finding of posterior diversity in MERS was seen quickly, that is, within 60 days of the first patient and in only 30 specimens. In this study of COVID-19 the cutoff date of the 7666 specimens was April 19, 2020 or approximately 140 days after the first documented case. The lack of posterior diversity in COVID-19 at a much later date than what was seen with MERS also argues against a non-laboratory source for this pandemic.

A useful avenue of future research for those working to find an animal source for COVID-19 would be new mathematical models or statistical methods that might find a "hidden" signal of posterior diversity in the current data set which shows none. And given access to the unprecedented quantity of human data for COVID-19 which can be mined via bioinformatics, efforts to find the "missing link" in the wild through search and sample should be a second priority to mining the human specimen data set.

SARS-CoV-1. A similar pattern of clinical cases that do not show a common ancestor in the human population but instead is evidence of posterior diversity is shown in the Text-Table on the left for SARS-CoV-1[67] compared to CoV-2 on the right[68]. SARS-CoV-1 shows clusters of cases in humans that are connected only by phylogenetic branches that reach back in time (all of the branches inside the purple box. This is because of the extensive mutational background created while being in the intermediate host, the civet. With CoV-2 on the right, every clinical case descends from the first clinical case, in the 19A clade. There are no background mutations to account for. I will show elsewhere that the first Clade A patient was at the PLA Hospital about 3 km from the WIV.

[67] https://pubmed.ncbi.nlm.nih.gov/14585636/
[68] https://nextstrain.org/

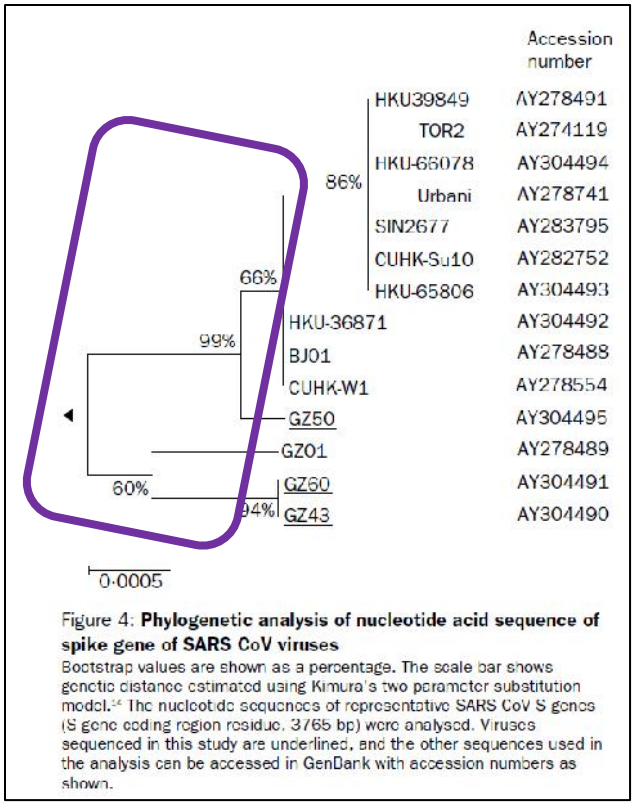
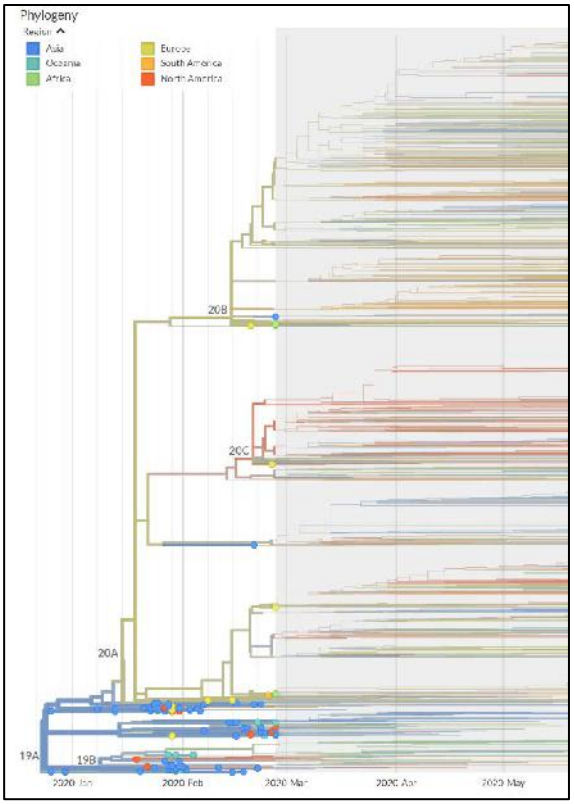
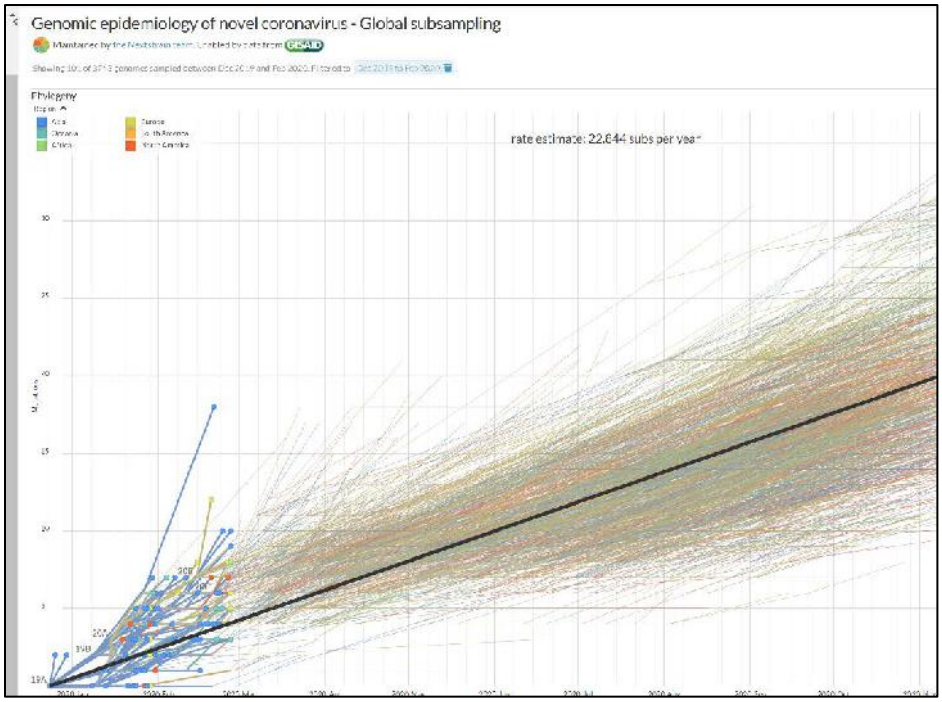


Figure 4: **Phylogenetic analysis of nucleotide acid sequence of spike gene of SARS CoV viruses**
Bootstrap values are shown as a percentage. The scale bar shows genetic distance estimated using Kimura's two parameter substitution model.[16] The nucleotide sequences of representative SARS CoV S genes (S gene coding region residue. 3765 bp) were analysed. Viruses sequenced in this study are underlined, and the other sequences used in the analysis can be accessed in GenBank with accession numbers as shown.



Given the rate of mutations of 22.8 per year for CoV-2 as shown in the Nextstrain graph below and a sequencing accuracy of about two calls per genome, CoV-2 could not have spent more than a few weeks in an intermediate host before a pattern of background mutations would be identified as posterior diversity. In the laboratory a pure culture on a single genome is used and the CoV-2 pattern is most consistent with a single pure culture infection a first human.

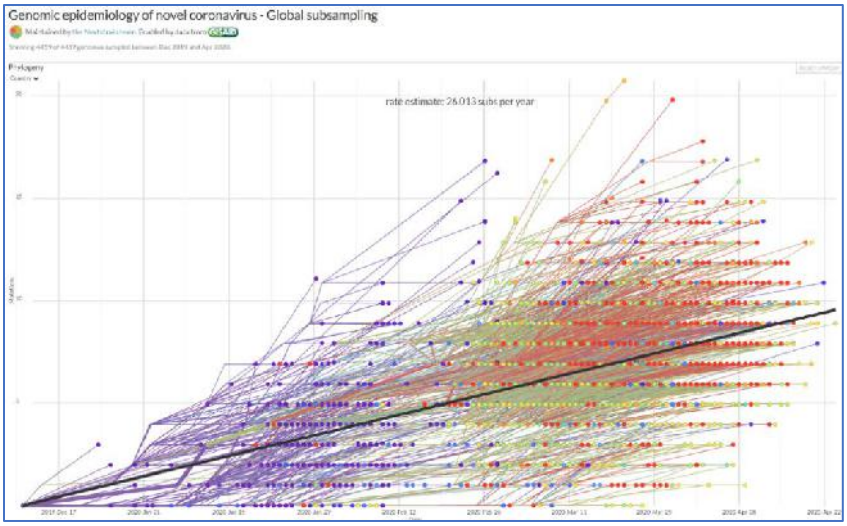


Non-zoonotic evolution. In a hypothetical in which there was a singular event in which one genetically pure virus infected one person and then the epidemic grow the development of the genetic diversity would have a clear, identifiable pattern: every new mutation would only appear on a background of the previous mutations.

The mutations in this virus are literally a personal tag. The general mutation rate leads to one mutation per patient. So, by definition, Patient Zero will have just one mutation. And then the 2-4 people that patient passes it to will have that mutation and then will add a new one, and so on. As time goes by two things happen: each patient gets a new mutation of their own and they pass on all the mutations of the past.

Since the virus has 29,900 nt and the mutation rate, as shown in this graph prepared by NextStrain is 26 mutations per year, there is very little chance a mutation will appear and then later get undone. By carefully going back in time, it is possible to literally name each person at each generation by the one (on average) new mutation they have and all of those that went before.

This graph of mutations on the Y-axis shows them gradually increasing and the color coding shows where they came from. In this infection, they only came from a previous patient and from the next previous patient and so on.



A NextStrain graphic.

How is that different from MERS, which was passed from camels to humans in a true zoonotic process?

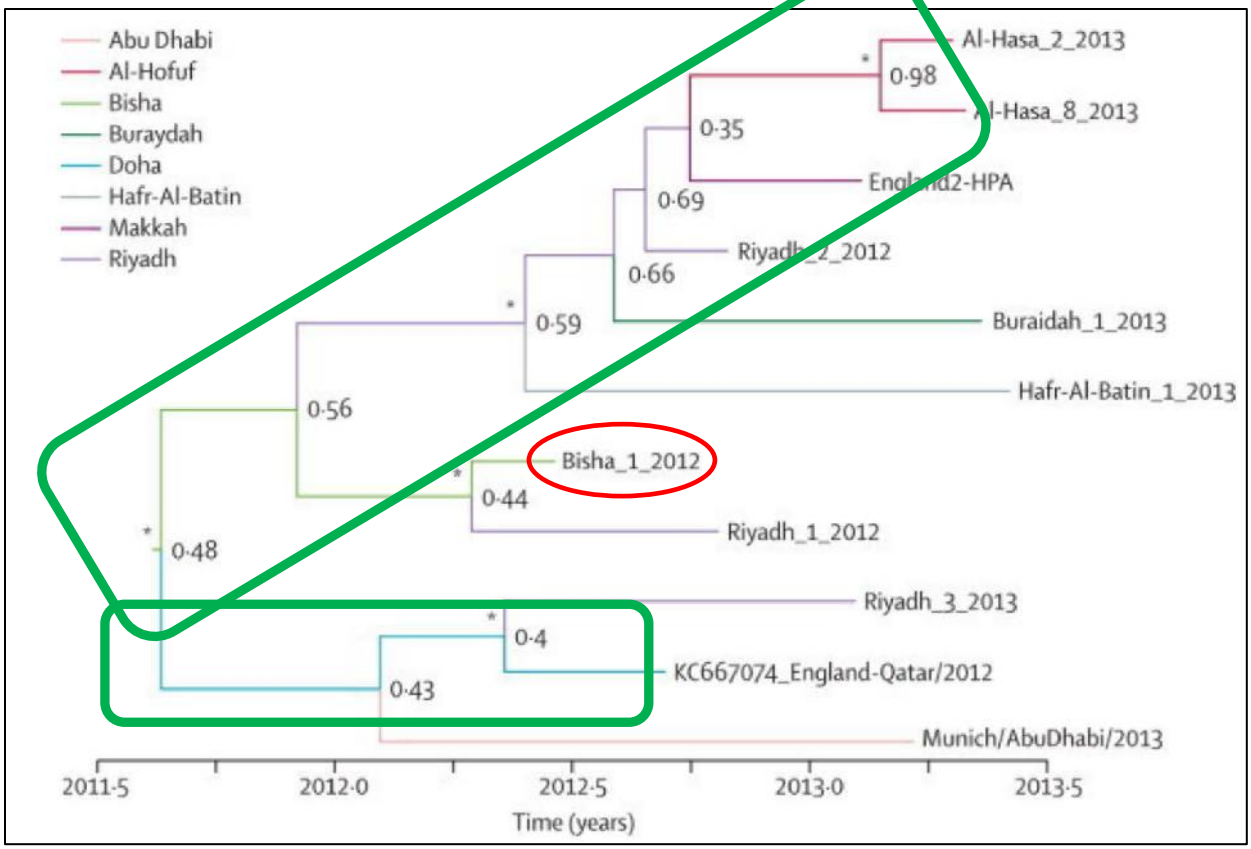
In a true zoonotic spread to humans there is usually an initiating species (in MERS it is bats), and then an intermediate species (in MERS it is camels), and then it moves to humans, either because of a new "enabling mutation" or for a non-domestic species, a chance encounter, and Source Zero and Patient Zero meet, and a cross species event occurs. But "Source Zero" doesn't stop there with one infection in one human; the virus also transmits itself vertically into the intermediate species. Source Zero also creates a vertical infection in the camels. Whether it is mild or not doesn't matter. The new human jumping gene is moving into a very diverse population of viruses, who have themselves been evolving since the first bat to camel transmission.

What is the outcome in terms of a test to show this is happening?

The diversity of the virus in humans becomes great, and the spots where the mutations occur don't match up to MERS Patient Zero like they do in COVID-19. In MERS, the virus in Patient Zero and the virus in a later infection are not direct descendants but cousins and only descended from an earlier virus that spent time in another camel population, collecting random mutations until it got the one it needed to infect humans, and then it begins again.

The chart below, from Lancet. 2013 Dec 14; 382(9909): 1993–2002, shows just how this works. The patient at Bisha is the earliest case in this chart (Patient Zero in the red circle). But notice, no other case comes from that patient. The viruses have such a diverse genetic background they appear to only be related to the Bisha virus with a posterior timeline of about one year. Their background is in the green boxes and it skips Patient Zero.

.



Even without knowing that camels are the zoonotic source for MERS, this data, from clinical sample only and without any field work in cave or camels, is all you need to know that this arose in the wild.

A paper just appeared with this analysis for a region of China and the posterior genomic diversity indicated a single starting point on December 1, 2019 for all cases. There was no posterior diversity. At this point with over 322,000 full genomes sequenced[69] and all showing an additive pattern of mutations and with none showing background diversity before the known appearance in Wuhan, the only conclusion is that there is no reservoir of genetic diversity.

On January 26, 2020 in an article in *Science* written by Jon Cohen, Kristian Andersen, an evolutionary biologist at the Scripps Research Institute who had analyzed sequences of CoV-2 to try to clarify its origin said: "The scenario of somebody being infected outside the market and then later bringing it to the market is one of the three scenarios we have considered that is still consistent with the data. It's entirely plausible given our current data and knowledge."

The negative predictive value of finding no posterior diversity in CoV-2 with 322,000 total infections sequenced, over 1000 in China, is 95%

Subjective Discount Factor: 95% (a one in 20 chance this is wrong)

[69] https://www.gisaid.org/

Below is the impact of the pack of posterior diversity on the likelihood of a zoonotic versus laboratory origin

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.809	0.191
Negative predictive value of lack of posterior diversity	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.809/9 = 0.085	
Normalize this step of analysis	0.085/(0.085 + 0.191) = 0.308	0.191/(0.085 + 0.191) = 0.692

Adjusted likelihood: Zoonotic origin (30.8%) and laboratory origin (69.2%)

Evidence: Opportunity.

The Wuhan Institute of Virology has publicly disclosed that by 2017 it had developed the techniques to collect novel coronaviruses, systematically modify the receptor binding domain to improve binding or alter zoonotic tropism and transmission, insert a furin site to permit human cell infection, make chimera and synthetic viruses, perform experiments in humanized mice, and optimize the ORF8 gene to increase human cell death (apoptosis).

Wuhan Institute of Virology scientists maps RBD and then takes a civet coronavirus that won't infect human cells, changes two amino acids in the receptor binding domain & it infects human cells.[70]

中国科技论文在线 http://www.paper.edu.cn

THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 280, No. 33, Issue of August 19, pp. 29588–29595, 2005

Printed in U.S.A.

Identification of Two Critical Amino Acid Residues of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Its Variation in Zoonotic Tropism Transition via a Double Substitution Strategy*

Received for publication, January 19, 2005, and in revised form, June 16, 2005
Published, JBC Papers in Press, June 24, 2005, DOI 10.1074/jbc.M500662200

Xiu-Xia Qu,[a,b] Pei Hao,[b,c] Xi-Jun Song,[a,b] Si-Ming Jiang,[a,b] Yan-Xia Liu,[a] Pei-Gang Wang,[a] Xi Rao,[a] Huai-Dong Song,[d] Sheng-Yue Wang,[e] Yu Zuo,[a] Ai-Hua Zheng,[a] Min Luo,[a] Hua-Lin Wang,[f] Fei Deng,[f] Han-Zhong Wang,[f] Zhi-Hong Hu,[f] Ming-Xiao Ding,[a] Guo-Ping Zhao,[c,e,g,h] and Hong-Kui Deng[a,i]

From the [a]Department of Cell Biology and Genetics, College of Life Sciences, Peking University, Beijing 100871, the [c]Bioinformation Center/Institute of Plant Physiology and Ecology/Health Science Center, Shanghai institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, the [d]State Key Laboratory for Medical Genomics/Pôle Sino-Francais de Recherche en Sciences du Vivant et Génomique, Ruijin Hospital Affiliated with the Shanghai Second Medical University, Shanghai 200025, the [e]Chinese National Human Genome Center, 250 Bi Bo Road, Zhang Jiang High Tech Park, Shanghai 201203, the [f]State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, 430071, and the [g]State Key Laboratory of Genetic Engineering/Department of Microbiology, School of Life Science, Fudan University, Shanghai 200433, China

Baric & Shi at WIV take bat coronavirus that won't infect human cells, change S746R to add an ARG at S1/S2 site to make furin-like cleavage site, & the new coronavirus infects human cells.[71]

Baric & Shi of WIV create completely synthetic coronavirus from bat spike & mouse adapted backbone that no treatment, monoclonal antibody, or vaccine will touch.[72]

- "Using the SARS-CoV reverse genetics system2, we generated and characterized a chimeric virus expressing the spike of bat coronavirus SHC014 in a mouse-adapted SARS-CoV backbone.

- The results indicate that group 2b viruses encoding the SHC014 spike in a wild-type backbone can efficiently use multiple orthologs of the SARS receptor human angiotensin

[70] http://www.paper.edu.cn/scholar/showpdf/NUT2kN0INTT0gxeQh
[71] https://jvi.asm.org/content/jvi/89/17/9119.full.pdf
[72] https://pubmed.ncbi.nlm.nih.gov/26552008/

converting enzyme II (ACE2), replicate efficiently in primary human airway cells and achieve in vitro titers equivalent to epidemic strains of SARS-CoV.

- Additionally, in vivo experiments demonstrate replication of the chimeric virus in **mouse lung with notable pathogenesis.**

- Evaluation of available SARS-based immune-therapeutic and prophylactic modalities revealed poor efficacy; both monoclonal antibody and vaccine approaches failed to neutralize and protect from infection with CoVs using the novel spike protein.

- On the basis of these findings, we **synthetically re-derived an infectious full-length** SHC014 recombinant virus and demonstrate robust viral replication both in vitro and in vivo."

This study was conducted, with permission, during the gain of function moratorium put in place by NIH in 2014:

"These studies were initiated before the US Government Deliberative Process Research Funding Pause on Selected Gain-of-Function Research Involving Influenza, MERS and SARS Viruses (http://www.phe.gov/s3/dualuse/Documents/gain-of-function.pdf). This paper has been reviewed by the funding agency, the NIH. Continuation of these studies was requested, and this has been approved by the NIH."

Drs. Daszak and Shi becomes world's expert on ORF8 induced apoptosis by CoVs in human cells (HeLa) & maximizing lethality.[73]

> The full-length ORF8 protein of SARS-CoV is a luminal endoplasmic reticulum (ER) membrane-associated protein that induces the activation of ATF6, an ER stress-regulated transcription factor that activates the transcription of ER chaperones involved in protein folding [35]. We amplified the ORF8 genes of Rf1, Rf4092 and WIV1, which represent three different genotypes of bat SARSr-CoV ORF8 (S3C Fig), and constructed the expression plasmids. All of the three ORF8 proteins transiently expressed in HeLa cells can stimulate the ATF6-dependent transcription. Among them, the WIV1 ORF8, which is highly divergent from the SARS-CoV ORF8, exhibited the strongest activation. The results indicate that the variants of bat SARSr-CoV ORF8 proteins may play a role in modulating ER stress by activating the ATF6 pathway. In addition, the ORF8a protein of SARS-CoV from the later phase has been demonstrated to induce apoptosis [28]. In this study, we have found that the ORF8a protein of the newly identified SARSr-CoV Rs4084, which contained an 8-aa insertion compared with the SARS-CoV ORF8a, significantly triggered apoptosis in 293T cells as well.

This paper also demonstrates the collection of 64 novel bat coronaviruses from caves in southern China, including Yunnan where Dr. Shi has said is the location of the bat ancestor of CoV-2.

This evidence is necessary for a laboratory origin hypothesis in which genetic manipulation to create CoV-2 is a precursor to a laboratory accident. However, it does not per se, provide

[73] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5708621/

increased weight in favor of a laboratory origin. It is however provided here to be a guide for the kinds of investigations to be conducted if access to the WIV records is ever provided.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (30.8%) and laboratory origin (69.2%)

Evidence and Motive for laboratory furin site insertion:

A key to infectivity of coronaviruses is the addition, in nature or the laboratory, of a furin cleavage site (FCS) at the S1/S2 junction of the Spike Protein.

Furin cleavage sites (FCS) have been widely understood to be important for many viral infections, including HIV, influenza, and others. It has also been widely understood before now that lineage B coronaviruses do not have FCS.

It was therefore surprising when an examination of SARS-CoV-2 Spike Protein found an insertion of a 12-nt, 4-AA sequence near the junction of the S1/S2 subunits which creates a furin site that is essential to human infectivity and transmission. As expected from previous work, no lineage B (sarbecovirus) coronavirus has this feature. This is the most difficult "molecular fingerprint" of SARS-CoV-2 to explain having been acquired in the wild and for that reason there are no even passingly feasible theories.

One database of whole genome sequences of 386 coronaviruses was devoid of furin cleavage sites.[74] Another database of 2956 genomes of sarbecovirus strains sequences shows that none have a furin site.[75] This is a highly significant finding with a probability that sarbecovirus has a furin site in the wild of one in about 985.[76]

It has been known since 1994 that viral glycoproteins can be cleaved by secreted proteases, including furin.[77] Even before that, in 1992, it was known the peptide sequence R-X-K/R-R in surface glycoproteins was required for avian influenza viruses of Serotype H7 pathogenesis.[78] The first paper using furin inhibitors to define a role for an FCS in coronavirus-cell fusion was published in 2004.[79]

Since that time, it has become common practice to insert FCS during laboratory gain-of-function experiments to increase infectivity. The following Text-Table illustrates the scope of just a few of the experiments conducted, with the hyperlink to the paper in column one.

URL for Paper	Title of Paper
One	Characterization of a panel of insertion mutants in human cytomegalovirus glycoprotein B.
Two	Insertion of the two cleavage sites of the respiratory syncytial virus fusion protein in Sendai virus fusion protein leads to enhanced cell-cell fusion and a decreased dependency on the HN attachment protein for activity.

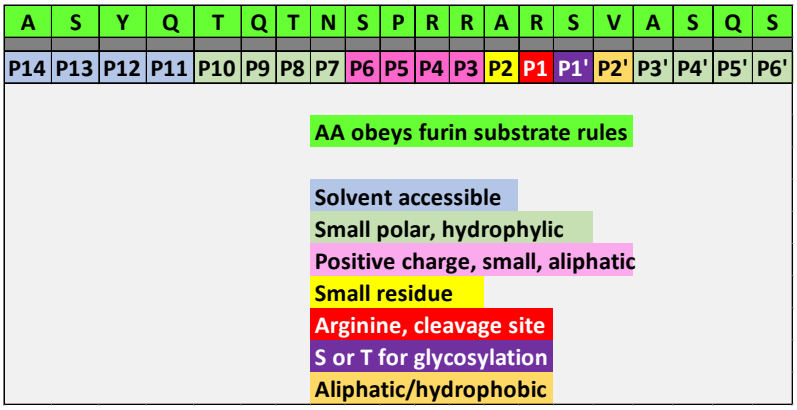
[74] https://academic.oup.com/bioinformatics/article/36/11/3552/5766118
[75] https://academic.oup.com/database/advance-article/doi/10.1093/database/baaa070/5909701
[76] When a series of samples are taken and none produce the result expected, the probability that this is a false negative finding can be estimated by taking the number of samples and dividing by three. Here, 2956 sarbecoviruses without a single furin site is a probability of one in 2956/3 or 985.
[77] https://www.ncbi.nlm.nih.gov/pubmed/8162439
[78] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7172898/pdf/main.pdf
[79] https://www.ncbi.nlm.nih.gov/pubmed/15141003

Three	Recombinant Sendai viruses expressing fusion proteins with two furin cleavage sites mimic the syncytial and receptor-independent infection properties of respiratory syncytial virus.
Four	Amino acid substitutions and an insertion in the spike glycoprotein extend the host range of the murine coronavirus MHV-A59
Five	Induction of IL-8 release in lung cells via activator protein-1 by recombinant baculovirus displaying severe acute respiratory syndrome-coronavirus spike proteins: identification of two functional regions.
Six	Coronaviruses as vectors: stability of foreign gene expression.
Seven	Experimental infection of a US spike-insertion deletion porcine epidemic diarrhea virus in conventional nursing piglets and cross-protection to the original US PEDV infection.
Eight	Minimum Determinants of Transmissible Gastroenteritis Virus Enteric Tropism Are Located in the N-Terminus of Spike Protein.
Nine	Reverse genetics with a full-length infectious cDNA of the Middle East respiratory syndrome coronavirus.
Ten	Construction of a non-infectious SARS coronavirus replicon for application in drug screening and analysis of viral protein function
Eleven	A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo.

The creation in the wild of a coronavirus FCS that is used as an example of what might have happened in SARS-CoV-2 is uninformative. In this case, a strain of influenza, in which a new polybasic site appears spontaneously leads to increased infectivity and lethality,[80] was reported by Tse *et al*. 2014. The mechanism of the FCS acquisition in this paper is an RNA polymerase dependent stuttering at a small, constrained loop in which one or more A nt were inserted, removing the strain in the loop and inserting an AAA codon which represents the basic amino acid lysine. No such method exists for the insertion of arginine, the amino acid in the CoV-2 furin site that needs to be created.

The insert generates a canonical 20 AA furin site sequence. In 2011 Tian et al.[81] published an analysis of 126 furin cleavage sites from three species: mammals, bacteria and viruses. The analysis showed that when the furin sites are recorded as a 20-residue motif, a canonical structure emerges. It includes one core cationic region (eight amino acids, P6–P2′) and two flanking solvent accessible regions (eight amino acids, P7–P14, and four amino acids, P3′–P6′).

[80] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3911587/
[81] https://www.nature.com/articles/srep00261



This figure above shows the 20-AA of the furin motif in SARS-CoV-2 (in green) with the P14 to P6' AA positions marked with the cleavage site being the amide bond between P1-R and the P1' residue. The motif is color coded with the requirements (in most cases, except for the positively charged AA requirements, most position requirements can be relaxed).

With the insertion, all 20 residues obey the rules as established by Tian. Since there are 20^4 different 4-AA peptides or 160,000 choices, it is remarkable that the 4 AA insert created a sequence that contained a small or cationic AA (8 AA/20 qualify), a cationic AA (3/20), another cationic AA (3/20), and a small AA (5/20) in that order. In fact, there are only 360 or the total or about 0.2% of all four amino acid inserts that would be expected to follow the exact rules for furin substrates. Of course, given the increase in infectivity SARS-CoV-2 has over other coronaviruses that do not have a well-designed furin cleavage site, selection pressure would drive this rare mutational event once it happened randomly. It would also be a likely choice for a laboratory designed furin cleavage site created *de novo*.

Based on the evidence that there are no furin cleavage sites in 2956 sarbecovirus (beta coronavirus) genome sequences[82], the likelihood that CoV-2 acquired the furin site from a wild sarbecovirus is one in 985 or 0.001. Because this is highly significant, we will use the conservative rule established in the beginning and use a likelihood of 0.05 for this evidence.

Subjective Discount Factor. 95% confidence (only a one in 20 chance this is wrong). Below is the calculation of the Bayesian adjustment.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.308	0.692
Negative predictive value of a lack of furin sites in sarbecovirus genomes	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.308/9 = 0.034	
Normalize this step of analysis	0.034/(0.034 + 0.692) = 0.047	0.692/(0.692 + 0.034) = 0.953

Adjusted likelihood. Zoonotic origin (4.7%), laboratory origin (95.3%).

[82] https://academic.oup.com/database/advance-article/doi/10.1093/database/baaa070/5909701

Evidence: Codon usage can distinguish insertion events in the wild from those created in the laboratory.

Not only is the insertion of an FCS peptide unique among lineage B coronaviruses, the nt sequence used for the process is more broadly unique among coronaviruses in general, regardless of lineage:

-CCT-**CGG-CGG**-GCA-

I will now use synonymous codon bias methods to try to inform the question of the origin of SARS-CoV-2.

Because of the redundancy of the genetic code, more than one 3-nt sequence specifies any given amino acid. For example, there are six codons that specify arginine, R. The frequencies with which such synonymous codons are used are unequal and have coevolved with the cell's translation machinery to avoid excessive use of suboptimal codons that often correspond to rare or otherwise disadvantaged tRNAs. This results in a phenomenon termed "synonymous codon bias," which varies greatly between evolutionarily distant species and possibly even between different tissues in the same species.

Decades of research has identified that all life forms, viruses, bacteria, and humans alike, use the codons in a signature pattern of frequency which can be used to identify a particular sequence of RNA or DNA as human or non-human; viral or non-viral.

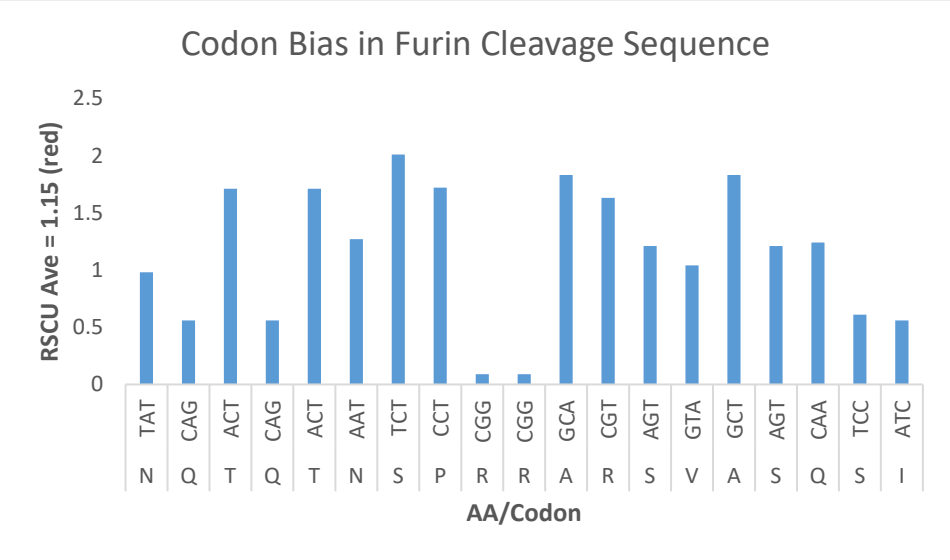
In this way, viruses in nature and scientists in the laboratory, with different goals and motivations, make distinguishing codon usage decisions which can sometimes provide a fingerprint of their source.

The Text-Table below contains the arginine codon usage for two populations, pooled data for SARS-CoV 2003 and related viruses and 13 Sars-CoV-2 human specimens from widely dispersed locations.

Codon	SARS-CoV 2003 and ten other evolutionary related viruses in the Nidovirales	SARS-CoV-2 from 13 Geo-locations
CGG	0.09	0.09
CGA	0.44	0.37
CGC	0.72	0.37
AGG	0.9	1.07
CGU	1.77	1.63
AGA	2.08	2.48

Since these values are of a type of multiplicative scale, they were fit using a log-normal distribution, which appears appropriate (although the sample size is small). Using the log mean and standard deviation and this distribution, the probability of finding a CGG codon is about 0.024. Assuming they are independent the probability of finding a CCG-CCG codon pair is effectively 0.024^2 or 0.00058. This is a likelihood of about one in 1700.

The following Figure shows the RSCU for the amino acids that comprise the new furin cleavage site in SARS-CoV-2. As one can see, the RSCU values are similar to each other with the exception of the RR dimer insert, which have a very low RSCU of 0.09.



The RSCU value for the CGG codon for R of 0.09 was taken from a 2004 paper of the RSCU for SARS-CoV 2003 and ten other evolutionary related viruses in the *Nidovirales* and is confirmed by 13 SARS-CoV-2 specimens obtained from diverse geographic locations. If one assumes that the RSCU observations are independent and that the probability distribution of these measurements is Gaussian (normal; a reasonable assumption), then one can calculate the probability of obtaining a result as small as 0.09. Removing the two 0.09 values, then the mean and standard deviation of the remaining values are 1.275 and 0.4992, respectively. Then the probability of a single 0.09 value is 0.0088. However, there are two 0.09 values. If we assume that these are independent findings, then the probability of both values being seen is 0.0088^2 or 7.7×10^{-5}. Using the RSCU of 0.2 from the Table above does not change the immense improbability of the usage of a CGGCGG codon pair in the wild.

Single Arginine CGG codon usage analysis suggests this will not be found in the wild.

The codon usage for SARS-CoV-2, like most coronaviruses studied, has a bias toward AT and away from GC nucleotides. The frequency of third position G use in CoV-2, for example, is 13%, 21%, 17%, and 16% for the spike protein, envelope, membrane, and nucleocapsid protein, respectively.

In that context, the scarcity of the CGG genome in SARS-CoV-2 and related coronaviruses, the relative synonymous codon usage, determined by the method of Behura and Severson,[83] was calculated and tabulated below. The color coding is blue for underutilized codons (RSCU < 1.0) and red for overutilized codons (RSCU > 1.0); light blue for RSCU values of 0.60 to 0.99 and

[83] https://www.ncbi.nlm.nih.gov/pubmed/22889422

light red for RSCU of 1.01 to 1.60. The highest RSCU usage of CGG is 1.21 in the membrane protein in the MERS virus but zero in SARS-CoV-2.

RSCU	SARS-CoV-2	Beta CoV Pangolin	SARS CoV	Bat SARS CoV	MERS CoV
Spike	0.29	0	0.19	0.08	0.25
Envelope	0	0	0	0	0
Membrane	0	0.35	0.74	0.24	1.21
Nucleocapsid	0.41	0.16	0.03	0.04	0.8

Looking at these five coronaviruses:

The largest structural protein of the coronaviruses is the spike protein, with 1273 amino acids. In SARS-CoV-2 there are 42 R residues, with only one RR dimer, the one in the insert that created SARS-CoV-2.

As a reminder none of these related coronaviruses have the 12-nucleotide insertion that forms the putative furin site in CoV-2. Interestingly, the pangolin coronavirus has no CGG residues in the spike protein. The significance of this is it makes the acquisition of this insert from pangolin by recombination impossible.

The smallest structural protein, the envelope protein, has 75 amino acids, including three R residues, but has no CGG codons in any of the related coronaviruses examined.

The SARS-CoV-2 membrane protein has 441 amino acids, 14 R residues and no CGG codons. Among related coronaviruses, this is the most unique finding of the four proteins for SARS-CoV-2 since the other four coronaviruses all utilize CGG to some extent in this protein. In the case of the MERS virus, this protein is the only occurrence in which this codon is overutilized.

The nucleocapsid protein has 418 amino acids and is responsible for packing the RNA genome. As expected for the role of R in protein-RNA interactions, it has 29 R residues and four RR dimers. None of the dimers use the CGGCGG sequence.

The nt usage of the 12-nt insert which forms the FCS cleavage site has a probability this sequence was selected for in the wild of one in 129,870.

A blast search was performed for the 12-nt inserted sequence and adjacent extensions and only the SARS-CoV-2 sequences were identified.

Shortening the search to just the two CGG-CGG codons was only slightly more fruitful. The Text-Table below shows the frequency of the middle half of the insert, CGGCGG, across the genomes of all seven known human coronaviruses, as well as a specimen bovine coronavirus and the bat and pangolin coronaviruses with greatest homology to SARS-CoV-2. Only a single example, outside of the Spike Protein gene, has been found.

Furin PBCS sequence	Beta Coronavirus	Total Arginine Dimers Anywhere	CGGCGG in Spike Protein *	CGGCGG Anywhere in genome *	CCGCCG Anywhere in genome
S**RR**K**RR**S	Human CoV-HKU1 GenBank: KF686346.1	12	0	0	0
K**RR**S**RR**A	Bovine CoV-Quebec GenBank: AF220295.1	12	0	0	0
PRRARSV	**SARS-CoV-2 Wuhan reference sequence GenBank: NC_045512.2**	**16**	**1; nt 23,606**	0	0
P**R**SV**R**S	MERS-CoV NCBI Reference Sequence: NC_019843.3	21	0	0	0
N**RR**S**R**GA	Human CoV-OC43 London/2011 GenBank: KU131570.1	16	0	0	0
None	Human CoV-229E GeneBank: KF514433.1	15	0	0	0
None	Human CoV NL63 NCBI Reference Sequence: NC_005831.2	9	0	0	0
None	SARS-CoV 2003 ZJ0301 from China GenBank: DQ182595.1	17	0	0	0
None	Bat coronavirus RaTG13 GeneBank: MN996532.1	11	0	1; nt 9394	0
None	Pangolin PCoV_GX-P4L GenBank: MT040333.1	10	0	0	0
	Total	139	1	0	0
* - Includes both in phase codons as well as out of phase, frameshift codons.					

To understand what this means for the search for the zoonotic source for SARS-CoV-2, a statistical approach was taken. Using the data from the nine viruses other than SARS-COV-2 there was a single incidence of the CGGCGG found in the bat coronavirus. Assuming 10,000 codons per genome, the frequency of CGGCGG in coronaviruses can be estimated at 2 per 45,000 codons or 4×10^{-5}. Therefore, the frequency of finding the center half of the SARS-CoV-2 insert is very small. This is consistent with the strong bias in all coronaviruses to place an A/U nt in the third codon position.

The last column above, the presence of -CCG-CCG- in these coronaviruses was included because it is the hybridization sequence partner for the negative strand sequence, which arises during genome replication. This eliminates the possibility of a strand jumping event to generate a CGGCGG codon dimer.

A similar analysis for the spike protein gene can be done. Since there are no instances of CGGCGG in the spike protein genome, and the gene is 3819 nucleotides long, there are 636 pairs of codons Thus, over the 9 other viruses, there are 5724 pairs of codons and no cases of the CGGCGG pair. To calculate the upper bound on the probability of such a pair from these data, one can use the Poisson “Rule of Three”, which yields a value of 3/5724 or 0.00052 with 95% confidence. Now examining the SARS-COV-2 genome, there was one instance of the pair in question out of 636 pairs. The probability of this happening if the true rate of this occurrence for a beta coronavirus is 0.00052 is 0.044. Obviously for smaller assumed rates of this occurrence, this would result in probabilities less than 0.044.

Since the 12-nt insert has been found nowhere in the coronavirus genomic universe, examining over 300,000 sequences and using the Poisson “Rule of Three” again, the upper bound on the frequency that it exists in nature is less than one in 100,000 with 95% confidence.

This observation in conjunction with the lack of finding the 12-nt sequence in any candidate zoonotic species makes unlikely a natural source for the virus. One line of investigation to establish a wild source for this infection would be to find a coronavirus strain with the 12-nt sequence somewhere in nature. The fact that 10 of the 12 nts are either G or C coupled, the documented bias against GC suggests this search would be futile.

Based on these analyses that demonstrate that the finding of a -CGG-CGG- codon pair in the furin site of CoV-2 is a highly improbable event, and using the conservative value of a one in 20 chance (the value for a p-value of 0.05), one can recalculate the likelihood of the choice between a zoonotic origin and a laboratory origin.

Subjective Discount Factor. 95% confidence (only a one in 20 chance this is wrong). Below is the calculation of the Bayesian adjustment.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin
Starting likelihood	0.047	0.953
Negative predictive value of the absence of the -CGG-CGG- pair in any coronavirus in nature	0.95	
Reduced by 95% Subjective Discount Factor	0.95 x 0.95 = 0.90	
Impact of this evidence	Reduces the likelihood of ZO by 90/10 or 9-fold. For every 100 tests, a true ZO would be seen 10 times and a non-ZO would be seen 90 times	
Impact of evidence calculation	0.047/9 = 0.005	
Normalize this step of analysis	0.005/(0.005 + 0.953) = 0.005	0.953/(0.953 + 0.005) = 0.995

Adjusted likelihood. Zoonotic origin (0.5%), laboratory origin (99.5%).

Evidence. Laboratory codon optimization uses CGG for laboratory insertions of arginine residues 50% of the time.

Codon optimization by recombinant methods (that is, to bring a gene's synonymous codon use into correspondence with the host cell's codon bias) has been widely used to improve cross-species expression of protein.

Though the opposite objective of reducing expression by intentional introduction of suboptimal synonymous codons has not been extensively investigated, isolated reports indicate that replacement of natural codons by rare codons can reduce the level of gene expression in different organisms. For example, one approach to vaccine development is to create an attenuated virus which comprises a modified viral genome containing nucleotide substitutions engineered in multiple locations in the genome, wherein the substitutions introduce synonymous de-optimized codons.

In US Patent 9,476,032[84] titled, "Attenuated viruses useful for vaccines," they state: "In one high-priority redesigned virus, most or all Arg codons are changed to CGC or **CGG** (the top two frequent human codons). This does not negatively affect translation." The patent contains numerous codon usages optimized for vaccine production, including the SARS-CoV virus, and in fact they use the CGG-CGG codon pair 45 times.

Beginning with a paper in 2004,[85] one motivation for codon-optimized SARS genomes is stated here: "The gene encoding the S protein of SARS-CoV contains many codons used infrequently in mammalian genes for efficiently expressed proteins. We therefore generated a codon-optimized form of the S-protein gene and compared its expression with the S-protein gene of the native viral sequence. S protein was readily detected in HEK293T cells transfected with a plasmid encoding the codon-optimized S protein."

Since that time, human optimized codons have been frequently used for coronavirus research, mostly in gain-of-function experiments. In that context the "molecular fingerprint" of CGG for R is one of those common laboratory reagent gene manipulators.

Other examples:

Examples of the use of CGG codon for arginine in coronavirus research	**Reference**
SARS was genetically modified to improve ACE2 binding using "human optimized" codons, like CGG for arginine, to grow better in the laboratory. The strains were more infective.Preparation of SARS-CoV S protein pseudotyped virus. "The full-length cDNA of	Wu, K. et al. Mechanisms of Host Receptor Adaptation by Severe Acute Respiratory Syndrome

[84] http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=9476032.PN.&OS=PN/9476032&RS=PN/9476032

[85] https://www.ncbi.nlm.nih.gov/pubmed/15367630

the SARS-CoV S gene was optimized according to human codon usage and cloned into the pCDNA3.1(+) vector (Invitrogen). The resulting "humanized" S sequence was identical with that of strain BJ01 at the amino acid level."	Coronavirus. J Biol Chem. 2012 Mar 16; 287(12): 8904–8911.
Predictions of future evolution of a virus are a difficult, if not completely impossible, task. However, our detailed structural analysis of the host receptor adaptation mutations in SARS-CoV RBD has allowed us to predict, design, and test optimized SARS-CoV RBDs that may resemble future evolved forms of the virus. "RBD might evolve into the human-optimized form by acquiring two mutations at the 442 and 472 position." SARS-CoV-2 acquired the mutation at position 472.	Fang Li. Receptor recognition and cross-species infections of SARS coronavirus. Antiviral Res. 2013 Oct; 100(1): 246–254.
Plasmid encoding a codon-optimized form of the SARS-CoV S protein of the TOR2 i	Wenhui Li, Chengsheng Z, et al., Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 2005 Apr 20; 24(8): 1634–1643.
The gene encoding the S protein of SARS-CoV contains many codons used infrequently in mammalian genes for efficiently expressed proteins. We therefore generated a codon-optimized form of the S-protein gene and compared its expression with the S-protein gene of the native viral sequence. S protein was readily detected in HEK293T cells transfected with a plasmid encoding the codon-optimized S protein (Fig. (Fig.1).1). No S protein was detected in cells transfected with a plasmid encoding the native S-protein gene.	Moore, MJ, Dorfman, T. Retroviruses Pseudotyped with the Severe Acute Respiratory Syndrome Coronavirus Spike Protein Efficiently Infect Cells Expressing Angiotensin-Converting Enzyme 2. J Virol. 2004 Oct; 78(19): 10628–10635.
Published in 2019 by **Dr. Zhengl-Li Shi,** entitled "Origin and evolution of pathogenic coronaviruses," reviews genetic optimized SARS viruses using human codons.	Cui, J, Fang, L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019; 17(3): 181–192.
In 2006, Montana scientists put a synthetic furin cleavage site into a SARS coronavirus by adding an R residue at position R667. They write: "We show that furin cleavage at the modified R667 position generates discrete S1 and S2 subunits and potentiates membrane fusion activity." Mutations were introduced by using	Follis, KE, York, J, Nunberg, JH. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell–cell fusion but does not affect virion entry. Virology 350 (2006) 358–369

QuikChange mutagenesis (Stratagene)[86]	
Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein is the title of a paper that shows that the expression of spike protein in vitro was greatly increased by expression cassette optimization.	Zhia, Y, Kobinger, GP, Jordan, H, et al. Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein
As for the human clec4C_1 and mouse clec14A, they showed very similar profiles with spike genes, especially with bat SARS-CoV, in the arginine coding groups, showing the high RSCU values over 2.50 in AGA.	Ahn,I, Jeong, B-J, Son, HS. Comparative study of synonymous codon usage variations between the nucleocapsid and spike genes of coronavirus, and C-type lectin domain genes of human and mouse. Experimental & Molecular Medicine volume 41, pages746–756, 2009.

One relevant paper,[87] in which arginine residues were being inserted into bovine herpesvirus-1, used primers to create RR dimers with nine separate -CGG-CGG- codon pairs. as testament to their broad use in the Wuhan Institute of Virology laboratory.

Scientists from the Wuhan Institute of Virology provided the scientific community with a technical bulletin on how to make genetic inserts in coronaviruses and proposed using the very tool that would insert this CGGCGG codon.

A Technical Appendix[88] entitled, "Detailed methods and primer sequences used in a study of genetically diverse filoviruses in Rousettus and Eonycteris spp. bats, China, 2009 and 2015, by Yang, Xinglou & Zhang, Yunzhi & Jiang, Ren-Di & Guo, Hua & Zhang, Wei & Li, Bei & Wang, Ning & Wang, Li & Rumberia, Cecilia & Zhou, Ji-Hua & Li, Shi-Yue & **Daszak, Peter** & Wang, Lin-Fa & **Shi, Zheng-Li.** (2017), from the Wuhan Institute of Virology identifies primer sequences for doing genetic experiments in coronaviruses and identifies CGG containing primers when a R amino acid is being inserted.

[86] Since the codon usage here was not reported I contacted Professor Nunberg to inquire which arginine codons were used. He replied: "Unfortunately, those files have all been archived and access to the nt sequences would involve considerable digging. If it is useful to you, I typically choose codons that are more frequent in highly expressed human proteins."
[87] From the Wuhan Institute of Virology; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7125963/
[88] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5382765/

Given that there are two codons of six possibilities that are used in codon optimization, CGG and CGC, the finding of a CGG pair would have a likelihood of happening by chance of (2/6) times (2/6) or one in nine.

Subjective Discount Factor: 80% (this has a probability of being wrong one in five times). This is arbitrary. The calculation to make this adjustment in likelihood is shown here:

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.005	0.995
This is the outcome expected 8 of 9 times if this is codon optimization		0.88
Reduced by 80% confidence		0.88 x 0.8 = 0.704
Impact of this evidence		Increases the likelihood of LO by 70.4 divided by 29.6 or 2.378.
Impact of evidence calculation		0.995 x 2.378 = 2.37
Normalize this step of analysis	0.005/(2.37 + 0.005) = 0.002	2.37/(0.005 + 2.37) = 0.998

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: SARS-CoV-2 Spike Protein is Highly Optimized for ACE2 Binding and Human Cell Infectivity, a Finding that is Inconsistent with Natural Selection but is Consistent with Laboratory Creation

Summary:

- Andersen et al.[89] hypothesized that if the CoV-2 interaction with the human ACE2 was apparently "not ideal," it was evidence that CoV-2 arose by natural selection.

- The alternative hypothesis would be that a finding that CoV-2 was optimized for ACE2 binding and human infection from the initial infection would be evidence of laboratory creation.

- Andersen relied on a paper for the "not ideal" interaction that relied on a computer algorithm rather than laboratory data, was qualitative in nature, sampled only five amino acids or 0.45% of the interaction region, and was over-interpreted.

- The analysis of the Baric et al. paper cited by Andersen as evidence the interaction was not ideal was reexamined, and it was concluded that Andersen had over-interpreted the paper. The paper was a computer simulation study of only 5 of 201 amino acids in the CoV-2-ACE2 interaction region. Only one of the five amino acids discussed was said to be inferior to the equivalent amino acid in SARS-CoV-1; the remainder were either positive or neutral with respect to binding.

- More recently, Baric has clarified his thoughts concerning the CoV-2 ACE2 receptor binding interaction. In a December 31, 2020 *New England Journal of Medicine* paper[57] he wrote: "Early zoonotic variants in the novel coronavirus SARS-CoV that emerged in 2003 affected the receptor-binding domain (RBD) of the spike protein and thereby enhanced virus docking and entry through the human angiotensin-converting–enzyme 2 (hACE2) receptor. **In contrast, the spike-protein RBD of early SARS-CoV-2 strains was shown to interact efficiently with hACE2 receptors early on**." [emphasis added.]

- A comprehensive, laboratory-based, and quantitative paper by Starr et al. of all 201 amino acids in the receptor binding region, not just five amino acids, was examined. Fully 99.6% of all of the possible 3819[90] amino acid substitutions were tested for their effect on CoV-2 binding to ACE2. Only 21 substitutions of the 3819 improved ACE2 binding. Therefore, CoV-2 has been optimized for human ACE2 binding in 99.45% of the possible amino acids in its Spike Protein interaction region.

[89] https://www.nature.com/articles/s41591-020-0820-9
[90] There are 201 amino acids in the residue 331 to 531 interaction region and so 201 times the 19 possible alternative amino acids not found in CoV-2 equals 3819.

- To support this finding, Starr also made an examination of 31,570 CoV-2 sequences from human infections, looking for the 21 substitutions that had been shown to improve CoV-2 binding in the above in vitro laboratory experiments. Among the 31, 570 CoV-2 cases, they failed to find even a single case in which there was an amino acid substitution that improved binding at the time of writing this analysis.[91]

- Based on Andersen's hypothesis and its alternative, SARS-CoV-2 is fully optimized for interaction with the human ACE2 receptor and was at the time of the first patient. There is no evidence of an evolving SP binding region, as was seen with SARS-CoV-1. This is consistent with a laboratory optimized coronavirus which entered the human population fully evolved.

Analysis

Quote from Andersen: "While the analyses above suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal (reference 7) and that the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding (references 7,11).

Thus, the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 that permits another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is not the product of purposeful manipulation."

The apparent **hypothesis** for the above conclusion is:

"If the SARS-CoV-2 (CoV-2) Spike Protein interaction with the ACE2 receptor is not maximized, then it is evidence that the interaction is the product of natural selection and not purposeful (laboratory) manipulation."

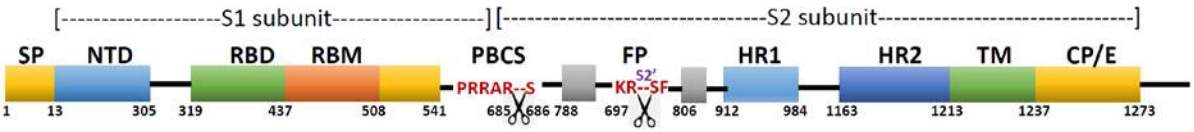
This would lead to an **alternative hypothesis**:

"If the CoV-2 Spike Protein interaction with the ACE2 receptor is maximized, then it is evidence that the interaction *was* the product of purposeful (laboratory) manipulation."

Background.

The Spike Protein (SP) structure and its functional domains are shown in this Figure. The S1 subunit is the initial host interaction portion while the S2 is the post-binding portion responsible for initiating host cell entry, with HR1, HR2, and TM being responsible for breaching the host cell membrane. Allowing viral RNA to enter the cell.

[91] The recent finding of the N501Y variant, first in the UK, and now spreading globally, is evidence of the power of this analysis. N501Y is one of only five potential substitutions in the Starr analysis that had a major effect in improving ACE2 binding.



The interaction of the SP portions which interact with the ACE2 of the host cell, which begins the internalization, infectious process, are contained in the Receptor Binding Domain (RBD) and to a lesser extent the Receptor Binding Motif (RBM), specifically residues 331 to 531. Herein, residues 331 to 531 are called the "interaction region."

Evidence given by Andersen:

Reference 7 in the Andersen paper above is a Ralph Baric paper[92] from early in the pandemic (submitted January 22, 2020) and examines five key residues in the receptor binding domain of the Spike Protein (SP) and whether they are "ideal" for interacting with the ACE2 of human cells. The entire paper is based on computer calculations or prior laboratory work but importantly does not do any new "wet" lab work with CoV-2.

Baric et al. had previously identified five amino acid residues that are important for SP-ACE2 interaction. Using the amino acid numbers of CoV-2, these amino acids are: 455, 486, 493, 494, and 501. Baric opines that the most critical residues are 493 and 501 and the next most important residues are 455, 486, and 494. The authors then discuss each amino acid in turn:

Residue 493: "Gln493 in 2019-nCoV RBD is compatible with hot spot 31, suggesting that 2019-nCoV is capable of recognizing human ACE2 and infecting human cells." In this analysis, 4 of the 20 amino acids are probed.

Residue 501: "This analysis suggests that 2019-nCoV recognizes human ACE2 less efficiently than human SARS-CoV (year 2002) but more efficiently than human SARS-CoV (year 2003). Hence, at least when considering the ACE2-RBD interactions, 2019-nCoV has gained some capability to transmit from human to human."

Direct binding evidence has shown that this statement is misleading, and CoV-2 binds the ACE2 receptor about ten-times better than SARS-CoV (year 2002).[93] In this analysis 3 of the 20 amino acids are probed.

Residues 455, 486, and 494: First, Baric et al. state: "Leu455 of 2019-nCoV RBD provides favorable interactions with hot spot 31, hence enhancing viral binding to human ACE2."

Next, they state: "Phe486 of 2019-nCoV RBD provides even more support for hot spot 31, hence also enhancing viral binding to human ACE2." Importantly, they also talk about their own laboratory work on an "optimized" receptor binding domain and state: "Leu472 of human and

[92] https://jvi.asm.org/content/94/7/e00127-20
[93] https://www.cell.com/action/showPdf?pii=S0092-8674%2820%2931003-5 ; https://www.nature.com/articles/s41586-020-2179-y ; https://www.sciencedirect.com/science/article/pii/S0092867420302622 ; https://science.sciencemag.org/content/367/6483/1260

civet SARS-CoV RBDs provides favorable support for hot spot 31 on human ACE2 through hydrophobic interactions with ACE2 residue Met82 and several other hydrophobic residues **(this residue has been mutated to Phe472 in the optimized RBD)**." [emphasis added.]

Finally, they state: Ser494 in 2019-nCoV RBD still provides positive support for hot spot 353, but the support is not as favorable as that provided by Asp480. Overall, Leu455, Phe486, and Ser494 of 2019-nCoV RBD support the idea that 2019-nCoV recognizes human ACE2 and infects human cells."

In this analysis they probe 3 of 20 amino acid residues for position 480, 4 of 20 for position 486, and 4 of 20 for position 442.

As shown in the Figure below from the Baric paper, the in vitro designed, optimized human SP (red arrow) had the amino acid residues F, F, N, D, and T at these five key residues. Since CoV-2 was identical in only one of these five it was not "optimal" and, according to Andersen, it therefore was not laboratory derived.

B

Virus	Year	442	472	479	480	487
SARS - human	2002	Y	L	N	D	T
SARS - civet	2002	Y	L	K	D	S
SARS - human/civet	2003	Y	P	N	G	S
SARS - civet	2005	Y	P	R	G	S
SARS - human	2008	F	F	N	D	S
Viral adaption to human ACE2		F > Y	F > L > P	N = R >>> K	D > G	T >>> S
Optimized - human	In vitro design	F	F	N	D	T
Viral adaptation to civet ACE2		Y > F	P = L > F	R > K = N	G > D	T > S
Optimized - civet	In vitro design	Y	P	R	G	T
SARS - bat	2013	S	F	N	D	N
2019-nCoV – human	2019	L (455)	F (486)	Q (493)	S (494)	N (501)

Conclusion from the above paper: by examining five amino acid residues of the 200 residues encompassing the interaction region, and calculating the expected interaction of a total of 18 of the 4000 possible residues or 0.45% of all possibilities, they conclude CoV-2 can infect human cells, but is not optimized to do so. This data was twisted by Andersen to show 'strong evidence' of natural selection.

An alternative and comprehensive analysis in another paper:[94]

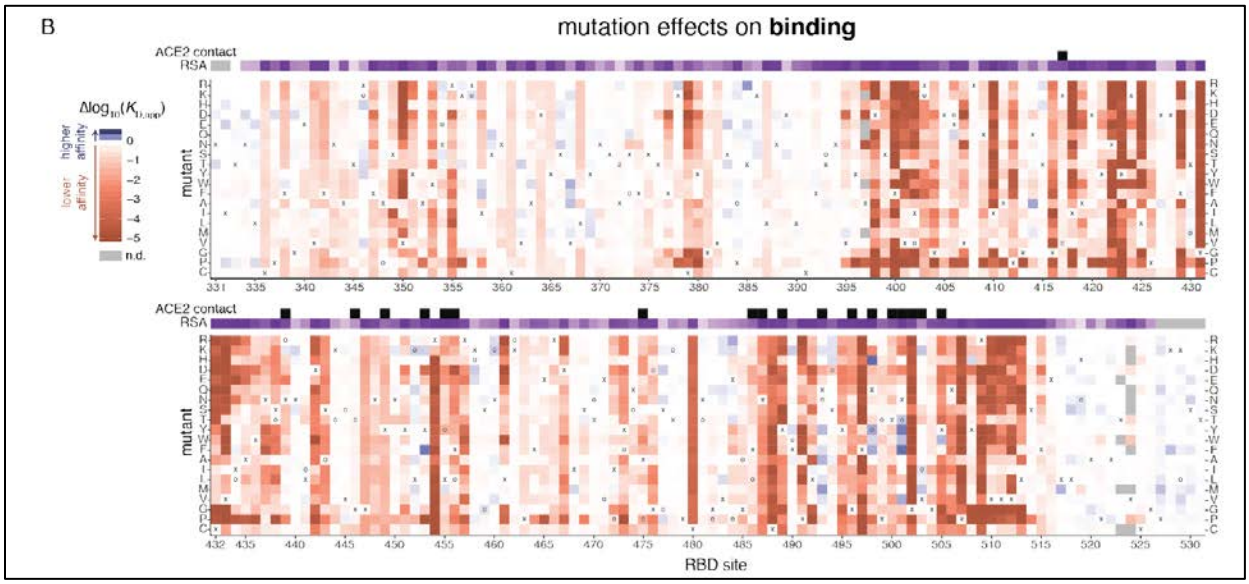
The receptor binding domain (RBD) of the CoV-2 SP is included in residues 331 to 531, a 201 amino acid sequence, of the SP. To examine the effect of each and every amino acid in each and every position, all 19 different amino acids were changed into all 201 positions of the RBD to the extent possible. Out of a total potential of 3819 different single amino acid variants, the scientists

[94] https://www.cell.com/action/showPdf?pii=S0092-8674%2820%2931003-5

were able to create 3804 of the potential variants or 99.6% of the possible variants. It is probable that the variants with the 0.4% amino acid substitutions could not be made for one reason or another. These 3804 were then tested for binding to the human ACE2. Finally, the RBD from SARS-CoV-1 also was tested.

The Figure below is the result of the experiment. Starting with amino acid 331 and ending with amino acid 531, the amino acids that were changed are in vertical columns and are color coded. Shades of brown are amino acid substitutions that reduce ACE2 binding affinity and blue are amino acid substitutions that improve binding, in all cases compared to the 'native' CoV-2 SP sequence. White is the color of a neutral substitution which neither enhances nor diminishes binding. Only the dark blue substitutions provide a strong improvement in ACE2 binding. There is a black square along the top row that denotes amino acids in the SP that interact with the ACE2 protein. Unlike in the Baric analysis above, in which only five amino acids were considered, this group of 19 amino acids provide a more complete interaction picture.

The first overarching observation is that most amino acid substitutions among the 201 amino acids are negative; while a large number are neutral. The fact that the vast majority of amino acid substitutions do not provide an improved ACE2 interaction is clear evidence that the CoV-2 SP interaction region is not newly evolved to the human ACE2 but arrived in the first patient having been "trained" to invade and kill human cells.

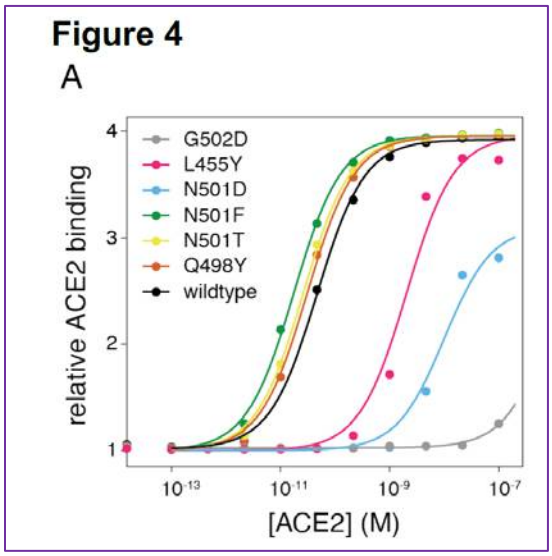


There are three levels of improved binding as designated by dark blue, medium blue, and pale blue. Out of the 3804 variants tested, there are 4 dark blue substitutions or 0.11% and 17 medium blue or 0.45%. According to the paper, the binding effect of the light blue could not be measured as different from the native sequence.

The conclusion of this comprehensive work is the demonstration that for 99.45% of the amino acids in the 201 amino acid interaction region, the CoV-2 choice is optimized, where any substitution is either detrimental or, at best, neutral with respect to the first step of CoV-2 entry to human cells, the binding step to the ACE2 receptor.

How much could CoV-2 binding be improved or made worse by substitutions during the human-to-human transmission of the pandemic?

The Figure 4 below, taken from the paper, shows that the three best amino acid substitutions have only a slight effect on the binding curve (Black is wildtype; curves to the left are better binding; curves to the right are worse binding). This is further evidence that CoV-2 is an optimized form of the original virus.



The authors also concluded that Anderson et al. was wrong: "An initially surprising feature of SARS-CoV-2 was that its RBD tightly binds ACE2 despite differing in sequence from SARS-CoV-1 at many residues that had been defined as important for ACE2 binding by that virus (Andersen et al., 2020; Wan et al., 2020)."

In fact, multiple studies have shown that CoV-2 binds ACE2 better than SARS-CoV-1, contradicting Andersen.

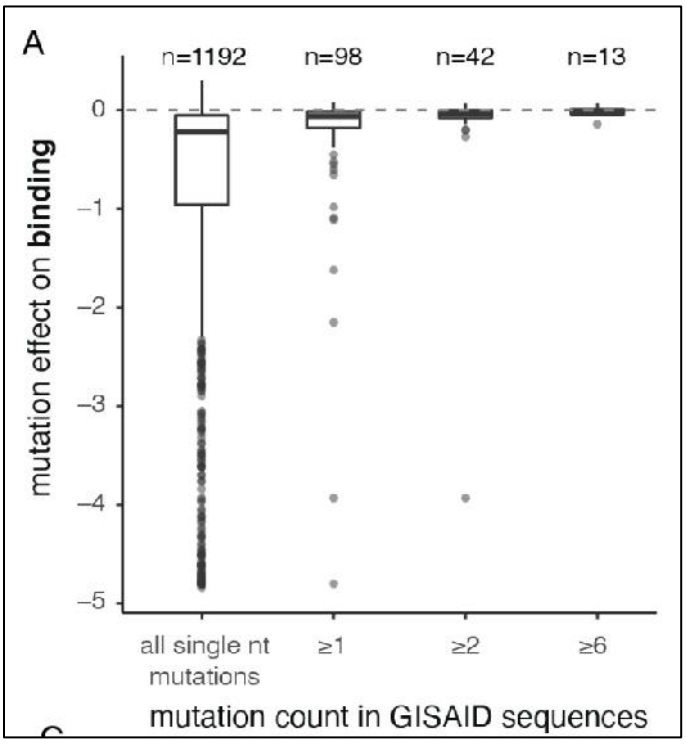
Is there evidence that CoV-2 in human circulation has mutations that enhance ACE2 binding?

Another measure of whether CoV-2 is optimized for human infection is to see if Spike Protein mutations have arisen during the pandemic that improve binding of the virus to the ACE2 receptor or if the SP amino acids are ideal from the very first human patient.

The Starr paper addressed this issue as well. A total of 31,570 human sequences were analyzed to see if any of the 21 amino acid substitutions from the binding experiments (or any other for that matter) were being selected for. That is, if there is any evidence of evolutionary pressure to improve SARS-CoV-2 infectivity.

Below is Figure 8 of the Starr paper. Of the 31,570 sequences, all mutations in the receptor interaction region were analyzed for their effect on ACE2 binding. The data below are for all examples of a single nt mutation (1192), two mutations (98), 3-5 mutations (42), and six or more (13) and the effect the mutation would have on ACE2 binding. The logarithmic scale has the

wildtype CoV-2 as 0 and each negative integer is a 10-fold reduction in affinity. Shockingly, there is not a single mutation that is above the 0 line, which would be an improved affinity for the ACE2 receptor. All of the mutations lower the receptor affinity.



Here are the results, in the words of Starr:

"Our discovery of multiple strong affinity-enhancing mutations to the SARS-CoV-2 RBD raises the question of whether positive selection will favor such mutations, since the relationship between receptor affinity and fitness can be complex for viruses that are well-adapted to their hosts (Callaway et al., 2018; Hensley et al., 2009; Lang et al., 2020). Strong affinity-enhancing mutations are accessible via single-nucleotide mutation from SARS-CoV-2 (Figure S8C), but **none are observed among circulating viral sequences in GISAID** (Figure 8A), and **there is no significant trend for actual observed mutations to enhance ACE2 affinity more than randomly drawn samples of all single nucleotide mutations** (see permutation tests in Figure S8D). **Taken together, we see no clear evidence of selection for stronger ACE2 binding, consistent with SARS-CoV-2 already possessing adequate ACE2 affinity at the beginning of the pandemic."** [emphasis added.]

It is striking that the authors, in observing the complete absence of any evidence for stronger ACE2 binding in over thirty thousand cases, would describe this as evidence of "adequate ACE2 affinity" and not as an exceptional finding of "optimized ACE2 affinity." Of course, calling the SP affinity exceptional from the beginning of the pandemic would beg the question of a laboratory derived virus.

Returning to the initial hypotheses, since the 3804 possible amino acids at the receptor interaction region of CoV-2 are 99.45% optimized for ACE2 binding, and there is not a single

example in 31,570 human CoV-2 genomes of a substitution that enhances ACE2 binding, the CoV-2 interaction with ACE-2 was maximized from the get-go.

Therefore, the hypothesis, "If the SARS-CoV-2 (CoV-2) Spike Protein interaction with the ACE2 receptor is not maximized, then it is evidence that the interaction is the product of natural selection and not purposeful (laboratory) manipulation," is **rejected**.

The alternative hypothesis, "If the CoV-2 Spike Protein interaction with the ACE2 receptor is maximized, then it is evidence that the interaction was the product of purposeful (laboratory) manipulation," is thus **accepted**.

At the time of this writing, a new RBD mutant N501Y has been observed. It is one of the five potential mutations that could be expected to increase RBD-ACE2 affinity.

This is the first example of evidence that will not be statistically quantified but treated as a 51%.49% preponderance of the evidence adjustment. The evidence is more consistent with having been optimized by various methods used in the laboratory than with the slow natural process as seen with SARS-CoV-1, and so the conservative rule that this is consistent with a laboratory origin (51%) versus zoonotic origin (49%) will be used. There will be no confidence adjustment.

The adjusted likelihoods are shown in the following table.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

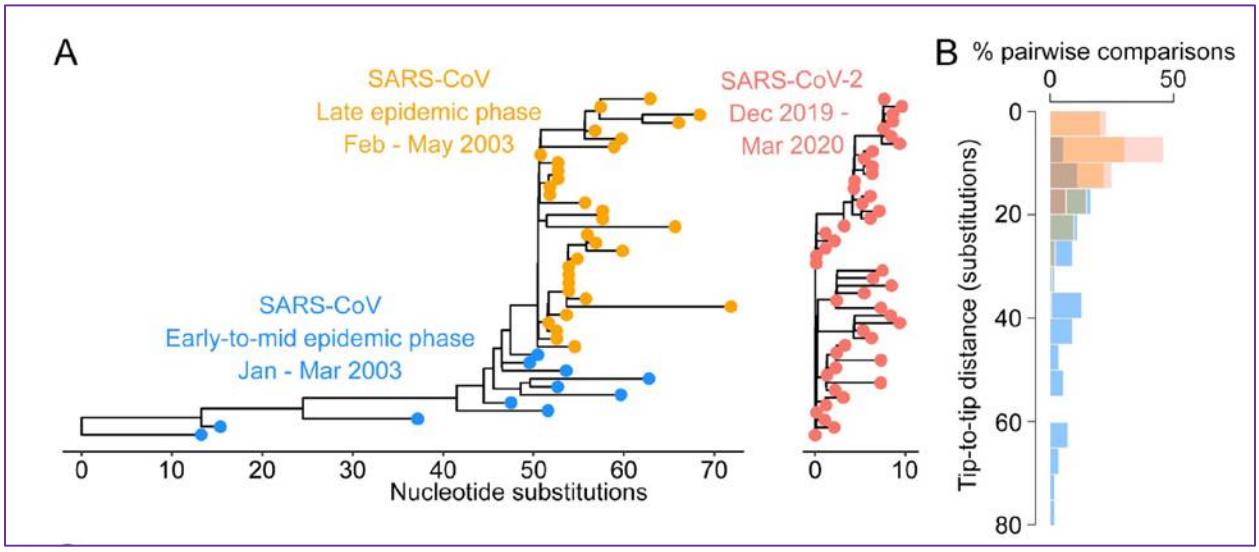
Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence. Whole genome comparison of human adaption of CoV-2 compared to SARS-CoV-1 is consistent with a "pre-adaption" of CoV-2 to the human host

A paper[95] entitled, "SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence?" by Shing Hei Zhan, Benjamin E. Deverman, and Yujia Alina Chan states in the abstract:

"In a side-by-side comparison of evolutionary dynamics between the 2019/2020 SARS-CoV-2 and the 2003 SARS-CoV, we were surprised to find that SARS-CoV-2 resembles SARS-CoV in the late phase of the 2003 epidemic, after SARS-CoV had developed several advantageous adaptations for human transmission. Our observations suggest that **by the time SARS-CoV-2 was first detected in late 2019, it was already pre-adapted to human transmission to an extent similar to late epidemic SARS-CoV**. **However, no precursors or branches of evolution stemming from a less human-adapted SARS-CoV-2-like virus have been detected.** The sudden appearance of a highly infectious SARS-CoV-2 presents a major cause for concern that should motivate stronger international efforts to identify the source and prevent re-emergence in the near future. [Emphasis added.]

The following Figure from the paper best illustrates the relative SNV adaption for SARS-CoV-1 versus CoV-2.



The paper also makes a tangential comment about posterior diversity: "It would be curious if no precursors or branches of SARS-CoV-2 evolution are discovered in humans or animals."

This is another example of evidence that will not be statistically quantified. The evidence is more consistent with having been adapted by various known methods used in a laboratory than with the slow natural process as seen with SARS-CoV-1, and so the conservative rule that this is consistent with a laboratory origin (51%) versus zoonotic origin (49%) will be used. There will be no confidence adjustment.

[95] https://www.biorxiv.org/content/10.1101/2020.05.01.073262v1

The adjusted likelihoods are shown in the following table.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

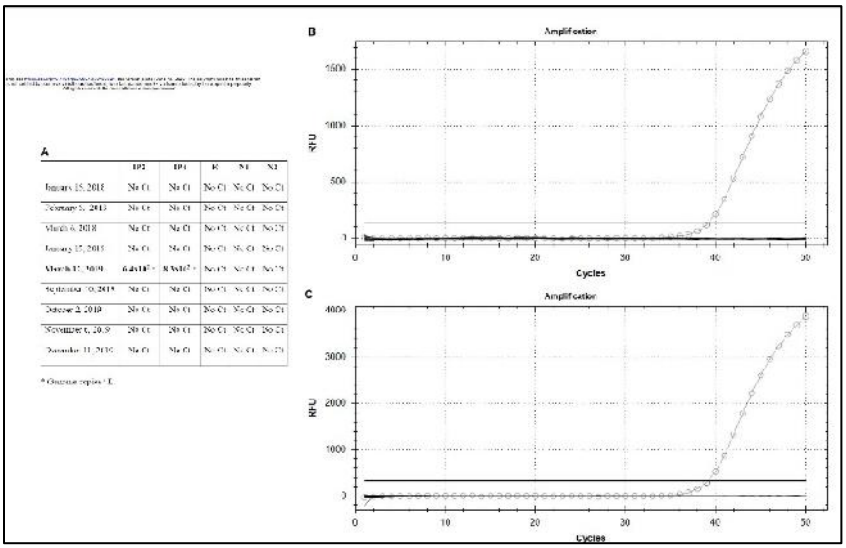
Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: Evidence of CoV-2 during early 2019 in wastewater from Barcelona, Spain is a false positive artifact

A paper entitled "Sentinel surveillance of SARS-CoV-2 in wastewater anticipates the occurrence of COVID-19 cases"[96] claims CoV-2 was present in Barcelona, Spain in March 2019. Specifically, they state:

"This possibility prompted us to analyze some archival WWTP samples from January 2018 to December 2019 (Figure 2). All samples came out to be negative for the presence of SARS-CoV-2 genomes with the exception of March 12, 2019, in which both IP2 and IP4 target assays were positive. This striking finding indicates circulation of the virus in Barcelona long before the report of any COVID-19 case worldwide."

This is a false positive



As shown above from the paper, they found 43/45 runs with zero and two runs had only 600-800 CoV-2 copies/L

But the limit of detection (LoD) of their assay is 1,000,000 CoV-2/L.

According to the Promega PCR assay FDA clearance package, the Ct at the LoD is 33-34 for the N1 and N2, respectively (Table 17, page 51).[97] Here the LoD is listed as 1 RNA/µL.

In the paper the Ct is 40 or 6-7 above the LoD.

This evidence is neutral as to origin and will not be used to adjust the likelihoods. It does reduce the credibility of some of the new origin theories coming out of China.

[96] https://www.medrxiv.org/content/10.1101/2020.06.13.20129627v1.full.pdf
[97] https://twitter.com/quay_dr/status/1340572543548227585/photo/1

Evidence: WHO and Dr. Shi have spoken of the singular nature of the beginning of COVID-19

On January 23, 2020 Dr. Shi wrote in the draft of her paper: "The almost identical sequences of this virus in different patients imply a probably recent introduction in humans…"[98] By February 3, 2020, when the final version of this paper was published, this sentence had been **deleted**.[99]

On April 23, 2020 the WHO stated: "All the published genetic sequences of SARS-CoV-2 isolated from human cases are very similar. This suggests that the start of the outbreak resulted from a single point introduction in the human population around the time that the virus was first reported in humans in Wuhan, China in December 2019."[100]

The evidence, like the lack of posterior diversity and seroconversion reported earlier, is more consistent with a single introduction in a laboratory accident. This evidence will not be used to adjust probabilities but is included because it could be a form of party admissions of unfavorable facts.

[98] RaTG13 paper as a preprint
[99] RaTG13 final Nature paper
[100] WHO document page 2 of 12

Evidence. As documented by Drs. Daszak, Humes, and Shi, mammalian biodiversity and bat species differences between Yunnan and Hubei Provence are significant and do not support a zoonotic origin

Summary. SARS-CoV-2 is most closely related to bat coronaviruses from Yunnan, a rural province in South West China. Wuhan, where the pandemic began, is a large urban city of 11 million inhabitants in north central China. These two areas are approximately 1900 km apart.

This is the US equivalent of the difference between New York City (population 8.4 million) and the Everglades in Florida, 2000 km away. The incongruent image of a bat or intermediate host in the Everglades somehow finding its way to New York City is a clear demonstration of the difficulty in this hypothetical transmission process. Nonetheless, a strict literature-based analysis will be conducted.

If COVID-19 is a zoonotic disease it must have travelled from bats to humans or from bats to an intermediate species to humans. Therefore, an examination of mammalian biodiversity differences and commonalities between Yunnan and Wuhan might provide useful information about the intermediate host or the particular bat species.

Peter Daszak, Zhengli-li Shi and colleagues published an August 2020 paper entitled, "Origin and cross-species transmission of bat coronaviruses in China,"[101] in which they make a number of observations that are relevant to this analysis. It should be remembered that both lead authors have made multiple, strong, public statements over many months where they assert that SARS-CoV-2 is a natural virus of zoonotic origin.

Yunnan and Hubei Provinces have very dissimilar mammalian diversity

Quoting from the Methods section of the Daszak, Shi paper:

"Defining zoogeographic regions in China:

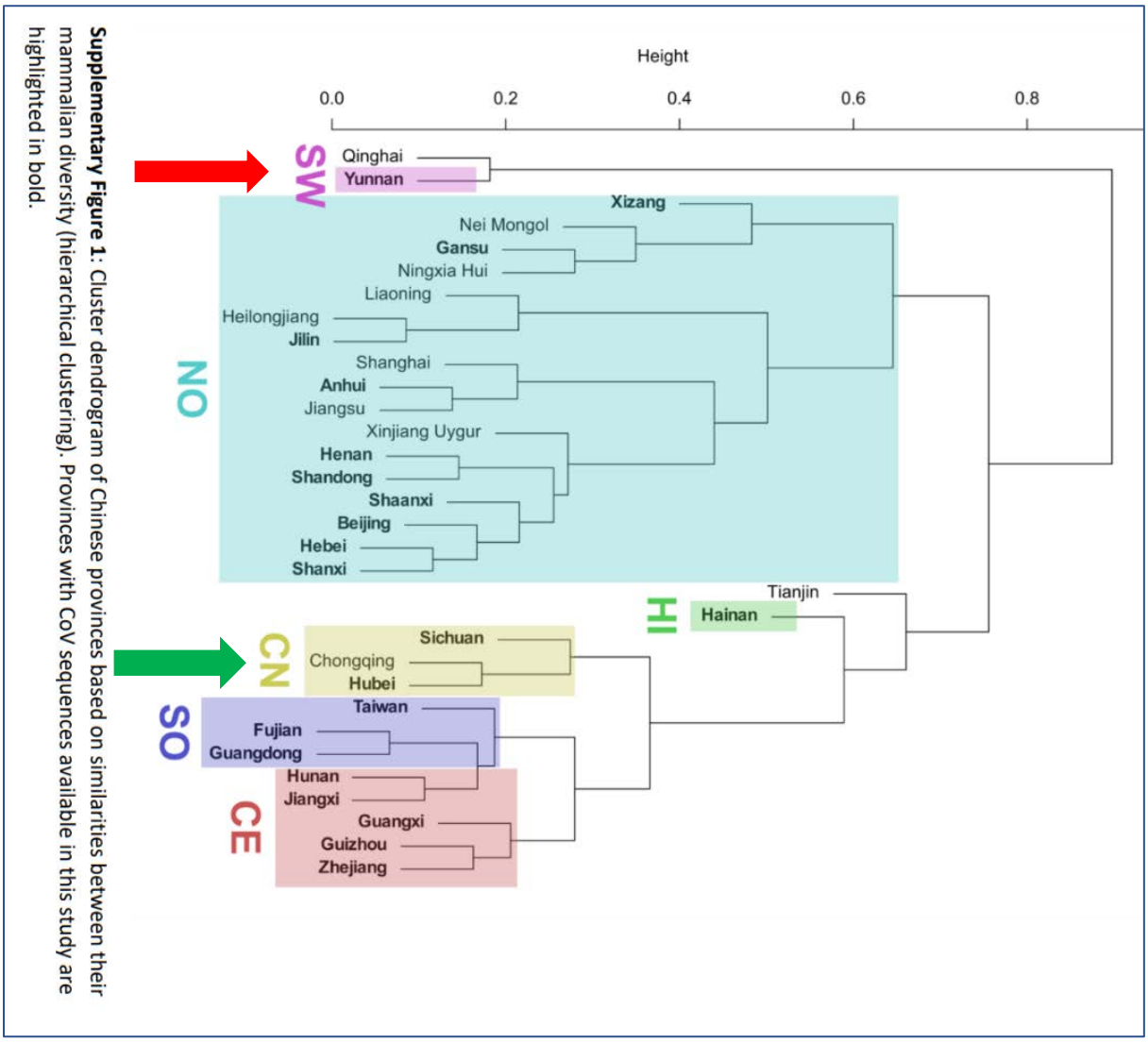
Hierarchical clustering was used to define zoogeographic regions within China by clustering provinces with similar mammalian diversity. Hierarchical cluster analysis classifies several objects into small groups based on similarities between them. To do this, we created a presence/absence matrix of all extant terrestrial mammals present in China using data from the IUCN spatial database and generated a cluster dendrogram using the function *hclust* with average method of the R package stats. Hong Kong and Macau were included within the neighboring Guangdong province. We then visually identified geographically contiguous clusters of provinces for which CoV sequences are available (Fig. 1 and Supplementary Fig. 1).

We identified six zoogeographic regions within China based on the similarity of the mammal community in these provinces: **SW (Yunnan province)**, NO (Xizang, Gansu, Jilin, Anhui, Henan, Shandong, Shaanxi, Hebei, and Shanxi provinces and Beijing municipality), **CN (Sichuan and Hubei provinces)**, CE (Guangxi, Guizhou, Hunan, Jiangxi, and Zhejiang provinces), SO (Guangdong and Fujian provinces, Hong Kong, Macau, and Taiwan), and HI.

[101] https://www.nature.com/articles/s41467-020-17687-3#Sec19

Hunan and Jiangxi, clustering with the SO provinces in our dendrogram, were included within the central region to create a geographically contiguous Central cluster (Supplementary Fig. 1). These six zoogeographic regions are very similar to the biogeographic regions traditionally recognized in China. The three β-CoV sequences from HI were included in the SO region to avoid creating a cluster with a very small number of sequences."

Below is a cluster dendrogram of Chinese provinces based on similarities between their mammalian diversity (hierarchical clustering). Provinces with CoV sequences available in this study are highlighted in bold.



Supplementary Figure 1: Cluster dendrogram of Chinese provinces based on similarities between their mammalian diversity (hierarchical clustering). Provinces with CoV sequences available in this study are highlighted in bold.

The y-axis height is a measure of the biodiversity with 1.0 being complete similarity and 0.0 being no similarity. As expected for the geography and location of the two provinces, Yunnan (red arrow above) and Hubei (green arrow above) have a height score of about 0.1, with seven branches and six nodes separating them. This is close to the biggest different in mammalian biodiversity of any two locations in all of China.

In conclusion, Daszak and Shi et al. demonstrate that the mammalian biodiversity between Yunnan and Hubei is very significant, reducing the options for a common intermediate host to be the natural conduit between bats and humans.

Shi, Humes, and Daszak statement: "SARS-CoV-2 is likely derived from a clade of viruses originating in horseshoe bats (*Rhinolophus* spp.). The geographic location of this origin appears to be Yunnan province."

This evidence will not be statistically quantified. The evidence reduces the biodiversity overlap needed to create a common intermediate species between the two provinces, and so the conservative rule that this is consistent with a laboratory origin (51%) versus zoonotic origin (49%) will be used. There will be no subjective discount factor adjustment.

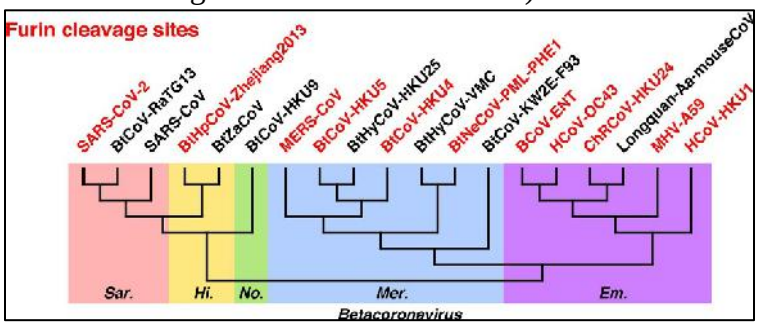
Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

Because of the rule on the use of significant figures, the likelihood does not change.

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

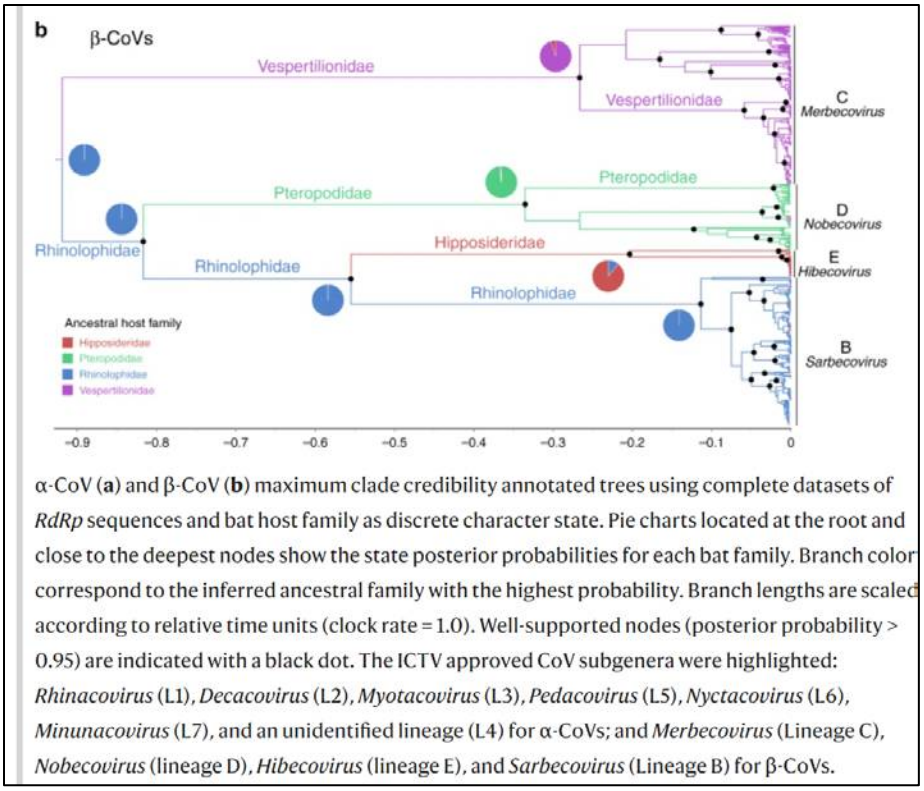
Evidence: The ancestor of SARS-CoV-2 can hypothetically only obtain a furin site by recombination outside of the sarbecovirus subgenera but there is strong evidence that coronavirus recombination is largely limited to the clade level, with limited evidence of sub-genera or genera recombination

- SARS-CoV-2 is a beta coronavirus, subgenera sarbecovirus and is the only sarbecovirus with a furin site.[102]
- Furin sites can be found in either alpha or gamma coronaviruses or the other beta coronavirus subgenera. The following Figure from reference 66 shows examples of such coronaviruses (furin containing viruses are shown in red):



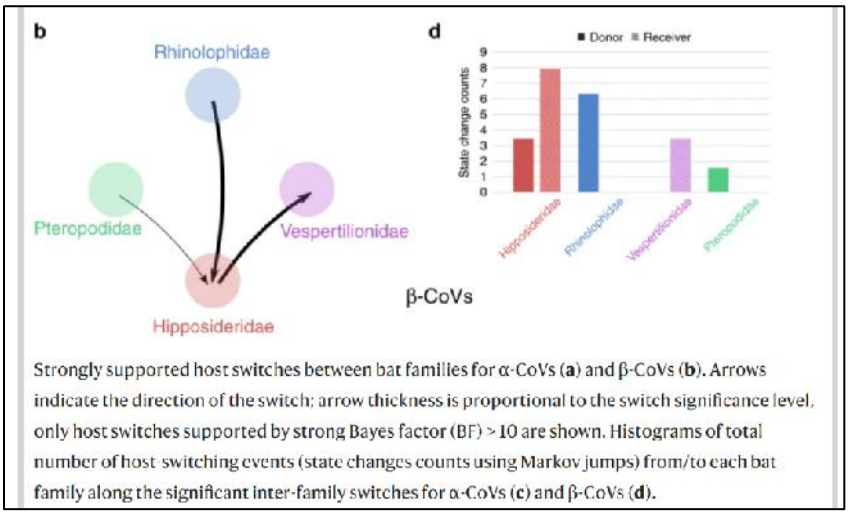
- To acquire a furin site in nature would require a co-infection between the CoV-2 sarbecovirus ancestor and a furin-containing non-sarbecovirus as shown above.
- However, there is no evidence of recombination in coronaviruses at either the genus level or the subgenus level; only at the clade level.[103][104]
- There is also evidence from Daszak and Shi that within the subgenera of the beta coronaviruses, there is bat host specificity. So, each subgenera of coronaviruses has a preferred bat host species. This reduces the opportunities for a co-host event to permit recombination.[105] The phylogeny below shows the problem of host incompatibility for beta coronaviruses:

[102] https://www.sciencedirect.com/science/article/pii/S1873506120304165#f0015
[103] file:///C:/Users/Steven%20Quay/Desktop/journal.pgen.1009272.pdf
[104] https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msaa281/5955840
[105] https://www.nature.com/articles/s41467-020-17687-3#Sec2



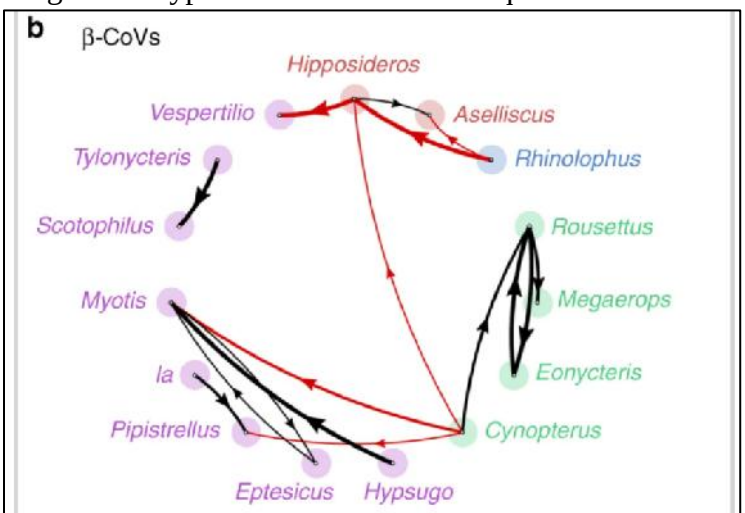
α-CoV (**a**) and β-CoV (**b**) maximum clade credibility annotated trees using complete datasets of *RdRp* sequences and bat host family as discrete character state. Pie charts located at the root and close to the deepest nodes show the state posterior probabilities for each bat family. Branch color correspond to the inferred ancestral family with the highest probability. Branch lengths are scaled according to relative time units (clock rate = 1.0). Well-supported nodes (posterior probability > 0.95) are indicated with a black dot. The ICTV approved CoV subgenera were highlighted: *Rhinacovirus* (L1), *Decacovirus* (L2), *Myotacovirus* (L3), *Pedacovirus* (L5), *Nyctacovirus* (L6), *Minunacovirus* (L7), and an unidentified lineage (L4) for α-CoVs; and *Merbecovirus* (Lineage C), *Nobecovirus* (lineage D), *Hibecovirus* (lineage E), and *Sarbecovirus* (Lineage B) for β-CoVs.

- Daszak and Shi also identified preferred directions of host switching. Since RaTG13, the closest coronavirus to SARS-CoV-2, is most closely related to viruses with bat hosts from the family, Rhinolophidae, it would be reasonable to expect furin-containing viruses from other bat hosts to migrate into Rhinolophidae, recombine by methods which have not been identified, and then the furin-containing sarbecovirus could evolve into the ancestor of SARS-CoV-2. Unexpectedly, Daszak et al. found host migration for the Rhinolophidae bats <u>only outward</u> and not inward, as required by the above, admittedly, convoluted process. The data Figure is shown here:



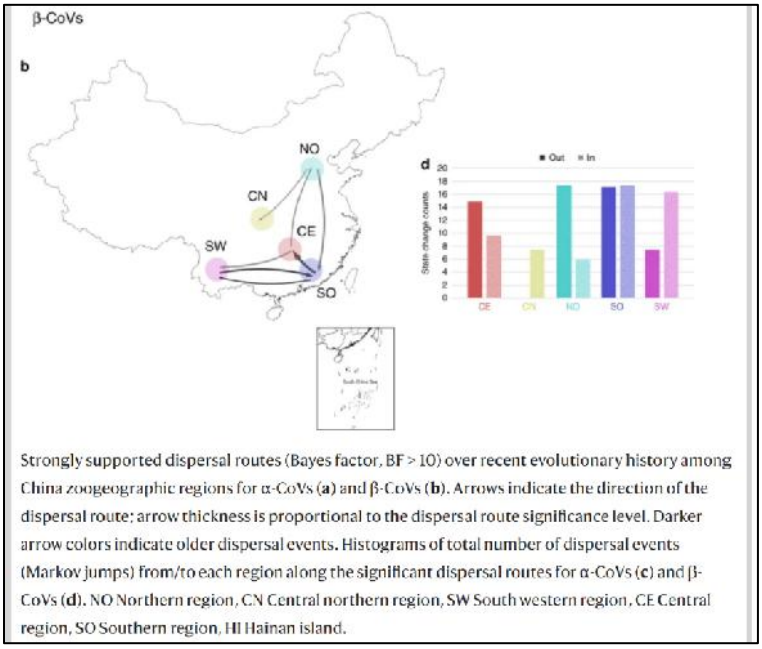
Strongly supported host switches between bat families for α-CoVs (**a**) and β-CoVs (**b**). Arrows indicate the direction of the switch; arrow thickness is proportional to the switch significance level, only host switches supported by strong Bayes factor (BF) > 10 are shown. Histograms of total number of host-switching events (state changes counts using Markov jumps) from/to each bat family along the significant inter-family switches for α-CoVs (**c**) and β-CoVs (**d**).

- Daszak and Shi also observed outward host switches from *Rhinolophus* at the genera level as well, also against a hypothesis for furin-site acquisition:



Strongly supported host switches between bat genera for α-CoVs (**a**) and β-CoVs (**b**) and their significance level (Bayes factor, BF). Only host switches supported by strong BF values >10 are shown. Line thickness is proportional to the switch significance level. Red lines correspond to host switches among bat genera belonging to different families, and black lines correspond to host switches among bat genera from the same family. Arrows indicate the direction of the switch. Genus names are colored according to the family they belong to using the same colors as in Figs. 2 and 3.

- Finally, this paper by Daszak and Shi states: "We used our Bayesian discrete phylogeographic model with zoogeographic regions as character states to reconstruct the spatiotemporal dynamics of CoV dispersal in China." If SARS-CoV-2 began in Yunnan and first crossed over into humans in Wuhan, this analysis should support a northernly spatiotemporal dispersal of beta coronaviruses. Unfortunately, Daszak and Shi cannot catch a break; their own data do not support the expected route of dispersion:



Strongly supported dispersal routes (Bayes factor, BF > 10) over recent evolutionary history among China zoogeographic regions for α-CoVs (**a**) and β-CoVs (**b**). Arrows indicate the direction of the dispersal route; arrow thickness is proportional to the dispersal route significance level. Darker arrow colors indicate older dispersal events. Histograms of total number of dispersal events (Markov jumps) from/to each region along the significant dispersal routes for α-CoVs (**c**) and β-CoVs (**d**). NO Northern region, CN Central northern region, SW South western region, CE Central region, SO Southern region, HI Hainan island.

As shown in the above Figure the only dispersal routes into Wuhan, which is in the CN region, are from the northern region. And the northern region has no inward dispersals from the SW, southwest region, where Yunnan and the origin of the ancestor of SARS-CoV-2, is located.

- Independent evidence documents that Hubei province does not have the bat species needed for SARS-CoV-2 reservoir host[106]

While statistical models of this data could be interesting and informative for general research about future spillovers, this is evidence will not be statistically quantified for this analysis. The evidence reduces the opportunities for subgenera co-infection and furin-site recombination into the CoV-2 ancestor and so the conservative rule that this is less consistent with a zoonotic origin (49%) versus laboratory origin (51%) will be used. There will be no subjective discount factor adjustment.

The results from the calculations are shown below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

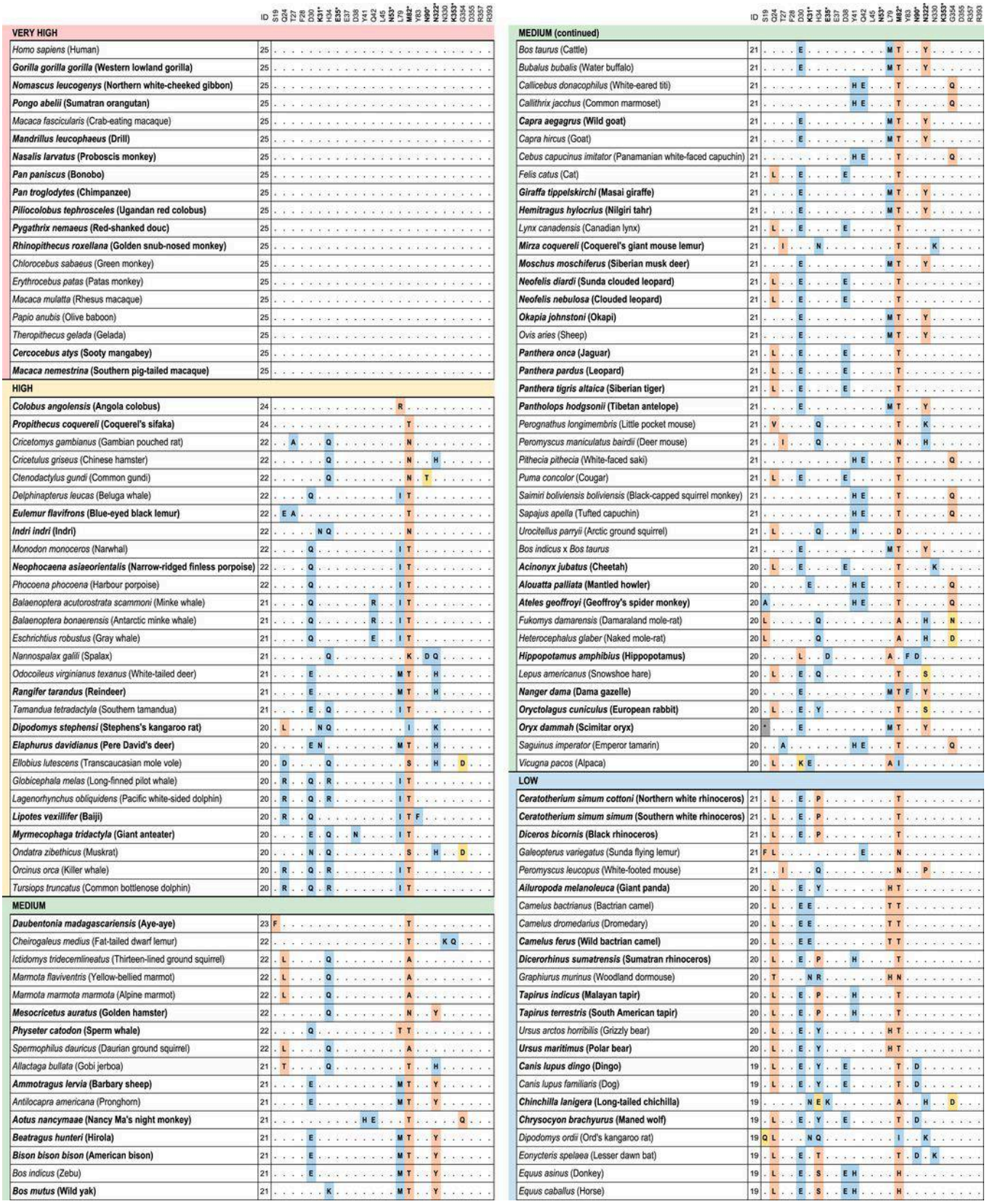
Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

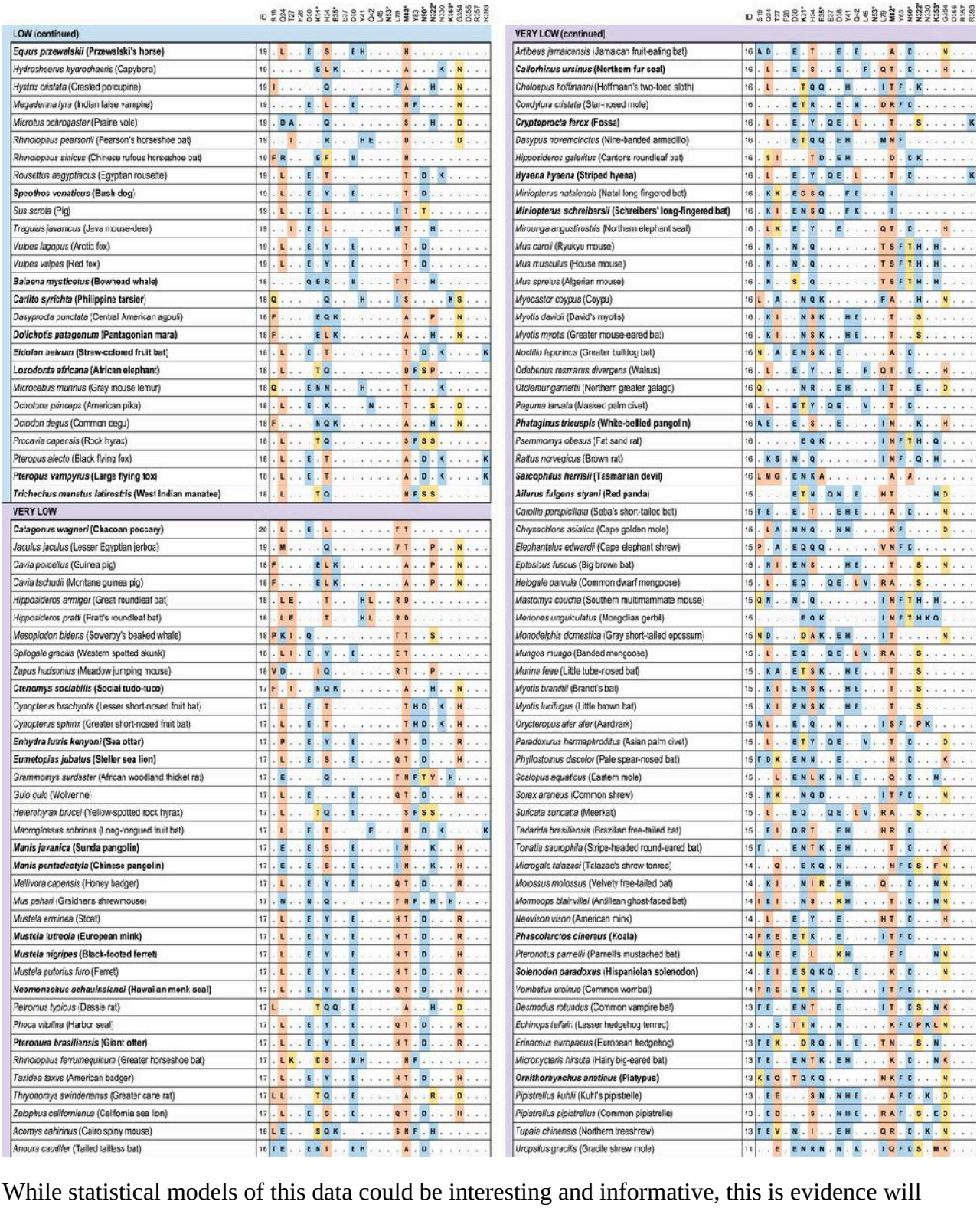
[106] file:///C:/Users/Steven%20Quay/Desktop/Zhangetal2009.pdf

Evidence: Of 410 vertebrate species tested for affinity to CoV-2 Spike Protein binding domain, primate ACE2 receptor, including human and VERO monkey cells, are the best at binding and bat species ACE2 are the worse, making direct bat-to-human host jumping extremely unlikely

- An examination of the ACE2 receptor binding domain amino acid sequences and their suitability for interacting with SARS-CoV-2 was performed in 410 vertebrates, including 252 mammals.[107]
- A five-category binding score was developed based on the conservation properties of 25 amino acids important for the binding between ACE2 and the SARS-CoV-2 spike protein.
- Only mammals fell into the medium to very high categories and only primates scored 25/25 for binding.
- This implies that SARS-CoV-2 is optimized for human ACE2-bearing cells from the first introduction into the human population, an observation that contradicts a zoonotic origin.
- It also suggests that other primates may be the proximate species from which SARS-CoV-2 entered the human population.
- Both VERO monkey kidney cells and ACE2 humanized mice would quality as an intermediate species by this criterion.
- Surprisingly, "all chiropterans (bats) scored low (n = 8) or very low (n = 29), including the Chinese rufous horseshoe bat, from which a coronavirus (SARSr-CoV ZC45) related to SARS-CoV-2 was identified."
- This is evidence that bats are probably not a reservoir host for SARS-CoV-2.
- A separate study observed: "Severe acute respiratory syndrome coronavirus 2 did not replicate efficiently in 13 bat cell lines."[108]
- The following two Tables are taken from the paper and are organized according to ACE2 SARS-CoV-2 affinity, from highest to lowest:

[107] https://www.pnas.org/content/117/36/22311
[108] https://wwwnc.cdc.gov/eid/article/26/12/20-2308_article


ID S19 Q24 T27 F28 D30 K31* H34 E35* E37 D38 Y41 Q42 L45 N53* L79 M82* Y83 N90* N322* N330 K353* G354 D355 R357 R393
VERY HIGH
Homo sapiens (Human) 25
Gorilla gorilla gorilla (Western lowland gorilla) 25
Nomascus leucogenys (Northern white-cheeked gibbon) 25
Pongo abelii (Sumatran orangutan) 25
Macaca fascicularis (Crab-eating macaque) 25
Mandrillus leucophaeus (Drill) 25
Nasalis larvatus (Proboscis monkey) 25
Pan paniscus (Bonobo) 25
Pan troglodytes (Chimpanzee) 25
Piliocolobus tephrosceles (Ugandan red colobus) 25
Pygathrix nemaeus (Red-shanked douc) 25
Rhinopithecus roxellana (Golden snub-nosed monkey) 25
Chlorocebus sabaeus (Green monkey) 25
Erythrocebus patas (Patas monkey) 25
Macaca mulatta (Rhesus macaque) 25
Papio anubis (Olive baboon) 25
Theropithecus gelada (Gelada) 25
Cercocebus atys (Sooty mangabey) 25
Macaca nemestrina (Southern pig-tailed macaque) 25
HIGH
Colobus angolensis (Angola colobus) 24
Propithecus coquereli (Coquerel's sifaka) 24
Cricetomys gambianus (Gambian pouched rat) 22
Cricetulus griseus (Chinese hamster) 22
Ctenodactylus gundi (Common gundi) 22
Delphinapterus leucas (Beluga whale) 22
Eulemur flavifrons (Blue-eyed black lemur) 22
Indri indri (Indri) 22
Monodon monoceros (Narwhal) 22
Neophocaena asiaeorientalis (Narrow-ridged finless porpoise) 22
Phocoena phocoena (Harbour porpoise) 22
Balaenoptera acutorostrata scammoni (Minke whale) 21
Balaenoptera bonaerensis (Antarctic minke whale) 21
Eschrichtius robustus (Gray whale) 21
Nannospalax galili (Spalax) 21
Odocoileus virginianus texanus (White-tailed deer) 21
Rangifer tarandus (Reindeer) 21
Tamandua tetradactyla (Southern tamandua) 21
Dipodomys stephensi (Stephens's kangaroo rat) 20
Elaphurus davidianus (Pere David's deer) 20
Ellobius lutescens (Transcaucasian mole vole) 20
Globicephala melas (Long-finned pilot whale) 20
Lagenorhynchus obliquidens (Pacific white-sided dolphin) 20
Lipotes vexillifer (Baiji) 20
Myrmecophaga tridactyla (Giant anteater) 20
Ondatra zibethicus (Muskrat) 20
Orcinus orca (Killer whale) 20
Tursiops truncatus (Common bottlenose dolphin) 20
MEDIUM
Daubentonia madagascariensis (Aye-aye) 23
Cheirogaleus medius (Fat-tailed dwarf lemur) 22
Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) 22
Marmota flaviventris (Yellow-bellied marmot) 22
Marmota marmota marmota (Alpine marmot) 22
Mesocricetus auratus (Golden hamster) 22
Physeter catodon (Sperm whale) 22
Spermophilus dauricus (Daurian ground squirrel) 22
Allactaga bullata (Gobi jerboa) 21
Ammotragus lervia (Barbary sheep) 21
Antilocapra americana (Pronghorn) 21
Aotus nancymaae (Nancy Ma's night monkey) 21
Beatragus hunteri (Hirola) 21
Bison bison bison (American bison) 21
Bos indicus (Zebu) 21
Bos mutus (Wild yak) 21
ID S19 Q24 T27 F28 D30 K31* H34 E35* E37 D38 Y41 Q42 L45 N53* L79 M82* Y83 N90* N322* N330 K353* G354 D355 R357 R393
MEDIUM (continued)
Bos taurus (Cattle) 21
Bubalus bubalis (Water buffalo) 21
Callicebus donacophilus (White-eared titi) 21
Callithrix jacchus (Common marmoset) 21
Capra aegagrus (Wild goat) 21
Capra hircus (Goat) 21
Cebus capucinus imitator (Panamanian white-faced capuchin) 21
Felis catus (Cat) 21
Giraffa tippelskirchi (Masai giraffe) 21
Hemitragus hylocrius (Nilgiri tahr) 21
Lynx canadensis (Canadian lynx) 21
Mirza coquereli (Coquerel's giant mouse lemur) 21
Moschus moschiferus (Siberian musk deer) 21
Neofelis diardi (Sunda clouded leopard) 21
Neofelis nebulosa (Clouded leopard) 21
Okapia johnstoni (Okapi) 21
Ovis aries (Sheep) 21
Panthera onca (Jaguar) 21
Panthera pardus (Leopard) 21
Panthera tigris altaica (Siberian tiger) 21
Pantholops hodgsonii (Tibetan antelope) 21
Perognathus longimembris (Little pocket mouse) 21
Peromyscus maniculatus bairdii (Deer mouse) 21
Pithecia pithecia (White-faced saki) 21
Puma concolor (Cougar) 21
Saimiri boliviensis boliviensis (Black-capped squirrel monkey) 21
Sapajus apella (Tufted capuchin) 21
Urocitellus parryii (Arctic ground squirrel) 21
Bos indicus x Bos taurus 21
Acinonyx jubatus (Cheetah) 20
Alouatta palliata (Mantled howler) 20
Ateles geoffroyi (Geoffroy's spider monkey) 20
Fukomys damarensis (Damaraland mole-rat) 20
Heterocephalus glaber (Naked mole-rat) 20
Hippopotamus amphibius (Hippopotamus) 20
Lepus americanus (Snowshoe hare) 20
Nanger dama (Dama gazelle) 20
Oryctolagus cuniculus (European rabbit) 20
Oryx dammah (Scimitar oryx) 20
Saguinus imperator (Emperor tamarin) 20
Vicugna pacos (Alpaca) 20
LOW
Ceratotherium simum cottoni (Northern white rhinoceros) 21
Ceratotherium simum simum (Southern white rhinoceros) 21
Diceros bicornis (Black rhinoceros) 21
Galeopterus variegatus (Sunda flying lemur) 21
Peromyscus leucopus (White-footed mouse) 21
Ailuropoda melanoleuca (Giant panda) 20
Camelus bactrianus (Bactrian camel) 20
Camelus dromedarius (Dromedary) 20
Camelus ferus (Wild bactrian camel) 20
Dicerorhinus sumatrensis (Sumatran rhinoceros) 20
Graphiurus murinus (Woodland dormouse) 20
Tapirus indicus (Malayan tapir) 20
Tapirus terrestris (South American tapir) 20
Ursus arctos horribilis (Grizzly bear) 20
Ursus maritimus (Polar bear) 20
Canis lupus dingo (Dingo) 19
Canis lupus familiaris (Dog) 19
Chinchilla lanigera (Long-tailed chichilla) 19
Chrysocyon brachyurus (Maned wolf) 19
Dipodomys ordii (Ord's kangaroo rat) 19
Eonycteris spelaea (Lesser dawn bat) 19
Equus asinus (Donkey) 19
Equus caballus (Horse) 19



While statistical models of this data could be interesting and informative, this is evidence will not be statistically quantified for this analysis. The evidence is another way of looking at the pre-adapted state of the CoV-2 for humans and suggests that primate animals, monkey cell cultures like the VERO cell, and humanized mice could be likely laboratory models that were used by the

WIV in GoF research. This will contribute a 51%/49% contribution in favor of laboratory compared to zoonotic origin. There will be no subjective discount factor adjustment.

The results from the calculations are shown below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: Did a Review of Samples Collected from a Mineshaft Cause the COVID-19 Pandemic?[109]

Abstract. The origin of the COVID-19 pandemic caused by SARS-CoV-2 has been hotly debated. Proponents of the natural spillover theory allege that the virus jumped species, possibly via an intermediary host, to cross over to humans via the wildlife trade or by other means. Proponents of a rival theory claim that the virus escaped from a laboratory in Wuhan. This research presents circumstantial evidence of a transmission route via a late 2019 review of samples collected from a mineshaft in Mojiang, Yunnan Province, China. It examines the activity at the Wuhan Institute of Virology in late 2019, when samples from a mineshaft associated with a suspected SARS outbreak were being reviewed. It proposes that spillover occurred during this review of samples including of a virus (BtCoV/4991) only 1% different to SARS-CoV-2 in its RNA-dependent RNA polymerase (RdRp).

It is a meticulous sourced analysis. It purposely avoids the question of whether SARS-CoV-2 was being grown or manipulated in the laboratory, but only addresses the evidence that events in the fall of 2019 are consistent with a laboratory accident.

This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[109] https://zenodo.org/record/4029545#.X-x_f9gzbOg. Author anonymous. A meticulously documented analysis that concludes an accident occurred at the Wuhan Institute of Virology during the fall of 2019. Includes many primary documents from Mandarin. No direct evidence of 'what' was the nature of the accident or if it was SARS-CoV-2.

Evidence: The Hunan market was not the source of SARS-CoV-2

From the WHO Terms of Reference for the investigation of the origin of SARS-CoV-2:[110]

"The Huanan wholesale market is a large market (653 stalls and more than 1180 employees) mainly supplying seafood products but also fresh fruits and vegetables, meat, and live animals. In late December 2019, 10 stall operators were trading live wild animals including chipmunks, foxes, racoons, wild boar, giant salamanders, hedgehogs, sika deer, and many others. Farmed, wild and domestic animals were also traded at the market including snakes, frogs, quails, bamboo rats, rabbits, crocodiles, and badgers. The market was closed on 1 January 2020, and several investigations followed, including environmental sampling, as well as sampling of frozen animal carcasses at the market. **Of the 336 samples collected from animals, none were PCR positive for SARS-CoV-2**, whereas 69 out of 842 environmental samples were positive by PCR for SARS-CoV-2. Sixty- one of those (88%) were from the western wing of the market. Of these, 22 samples were from 8 different drains and sewage, and 3 viruses were isolated, sequenced and shared on GISAID. These were virtually identical to the patient samples collected at the same time (>99.9 % homology)."

For contrast, with SARS-CoV-1 91 civets & 15 raccoon dogs in wet markets were tested with 106/106, 100% positive.[111]

This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[110] https://drive.google.com/file/d/1rx0W2efbE0R1Aq-lALWTqD22VsWbTlO-/view
[111] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1212604/

Evidence: Analysis of the hospital of admission for COVID-19 patients during December 2019 places "ground zero" for the outbreak somewhere along Line 2 of the Wuhan Metro System.

Line 2 carries one million people per day and services the Wuhan Institute of Virology, the Hunan Seafood Market, the high-speed rail system, and the Wuhan International Airport

A preprint manuscript[112] reported that the earliest genomic cluster of SARS-CoV-2 patients is a group of four individuals associated with the General Hospital of Central Theater Command of People's Liberation Army (PLA) of China in Wuhan. This cluster contains the "Founder Patients" of both Clade A and Clade B, from which every SARS-CoV-2 coronavirus that has infected every patient with COVID-19 anywhere in the world has arisen.

The PLA Hospital is about one mile from the Wuhan Institute of Virology (WIV) and the closest hospital to WIV. Both the PLA Hospital and WIV are serviced by Line 2 of the Wuhan Metro System. The Hunan Seafood Market is also located adjacent to Line 2. All patients between December 1st, 2019 and early January 2020 were first seen at hospitals that also are serviced by Line 2 of the Metro system.

With 40 hospitals located near seven of the nine Metro Lines, the likelihood that all early patients were seen at hospitals only near Line 2 by chance is about 1 in 68,500 (p-value = 0.0000146). The inference then would be that the early spread of SARS-CoV-2 was through human-to human transmission on Line 2.

Line 2 carries one million passengers per day and assuming most are round trip business workers going to and from work in the morning and evening, represents 500,000 riders or about 5% of the Wuhan population. A very recent publication determined that, in fact, 500,000 residents of Wuhan contracted COVID-19, a ten-fold upper estimate.[113] The coincidence of my prediction that 500,000 riders on Line 2 were likely exposed to SARS-CoV-2 in late 2019 and the recent admission from Chinese CDC that Wuhan had 500,000 COVID-19 cases is duly noted!

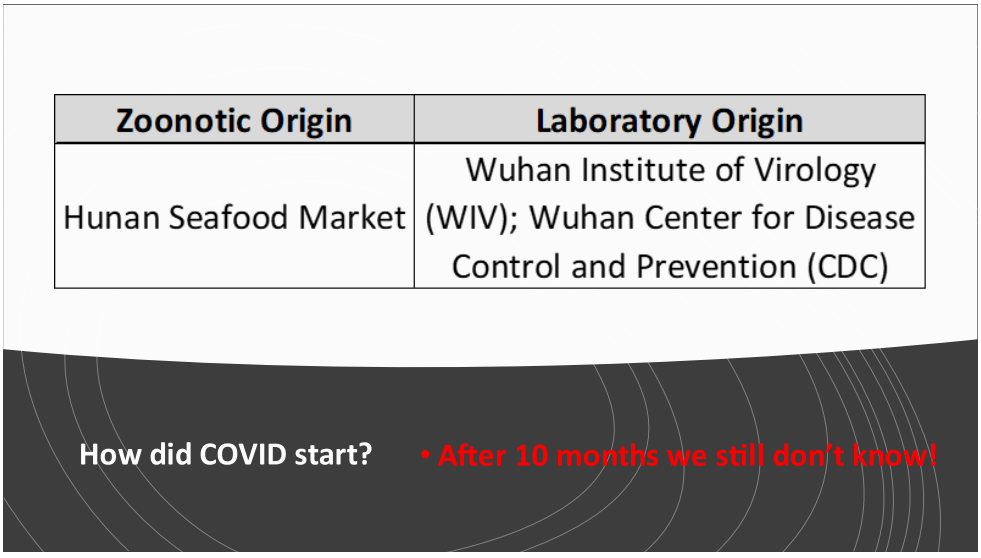
Line 2 connects to all eight other lines of the Wuhan Metro System (1, 3, 4, 6, 7, 8, 11, and Yanglu) facilitating rapid spread in Wuhan and Hubei Province, and also services both the high-speed rail station (Hankou Railway Station), facilitating rapid spread throughout China, and the Wuhan International Airport (Tianhe International Airport), facilitating rapid spread throughout Asia, Europe, and to the United States. In fact, direct human-to-human spread from the Reference Sequence patient to patients around the world is suggested by an unexpectedly reduced genome base substitution rate seen in patient specimens in cities with direct flights from Wuhan.

[112] https://zenodo.org/record/4119263#.X-rszNgzbOg
[113] https://mp.weixin.qq.com/s/LXTfDmsQLf3qZnu_S_MxcA ; https://thehill.com/policy/international/china/531935-study-shows-wuhan-coronavirus-cases-may-have-been-10-times-higher

In a separate paper by Quay and Dr. Martin Lee, Adjunct Professor of Statistics, UCLA, from May 2020, now accepted for publication in *Epidemics*,[114] the authors provide evidence that COVID-19 was appearing in California as early as the first week of 2020. This is likely due to direct flights connecting Line 2 to the Wuhan airport and then to San Francisco.

In conclusion, Line 2 of the Wuhan Metro System services the PLA Hospital with the first genomic cluster of patients with COVID-19, the hospitals where patients first went in December 2019 and early January 2020 and is the likely conduit for human-to-human spread throughout Wuhan, China, and the world.

The following slide overview provides a visual analysis of this evidence:



[114] https://www.researchgate.net/publication/341742303_COVID-19_May_Have_Have_Reached_United_States_in_January_2020_05272020

SARS 2003 and MERS 2015

- **Within four months (SARS 2003) and nine months (MERS 2015) we knew the zoonotic host**

GISAID

Virus detail

Virus name:	hCoV-19/Wuhan/WH01/2019
Accession ID:	EPI_ISL_406798
Type:	betacoronavirus
Lineage /GISAID Clade:	B (2)
Passage details/history:	Original

Sample Information

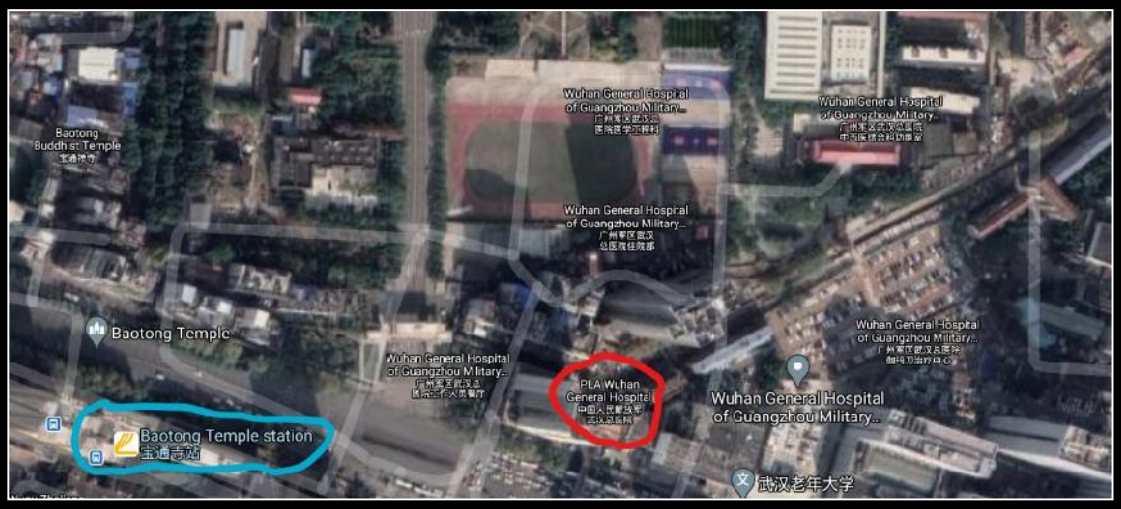
Collection date:	2019-12-26
Location:	Asia / China / Hubei / Wuhan
Host:	Human
Additional location information:	
Gender:	male
Patient age:	44
Patient status:	unknown
Specimen source:	Bronchoalveolar Lavage NCIT:C51913
Additional host information:	
Outbreak:	
Last vaccinated:	
Treatment:	
Sequencing technology:	DNBSEQ
Assembly method:	SPAdes v.3.12.0
Coverage:	83x
Comment:	

Institute Information

Originating lab:	General Hospital of Central Theater Command of People's Liberation Army of China
Address:	NO.627 Wuhan Road, Wuchang District, Wuhan, China

GISAID Database

Earliest cases at the PLA Hospital



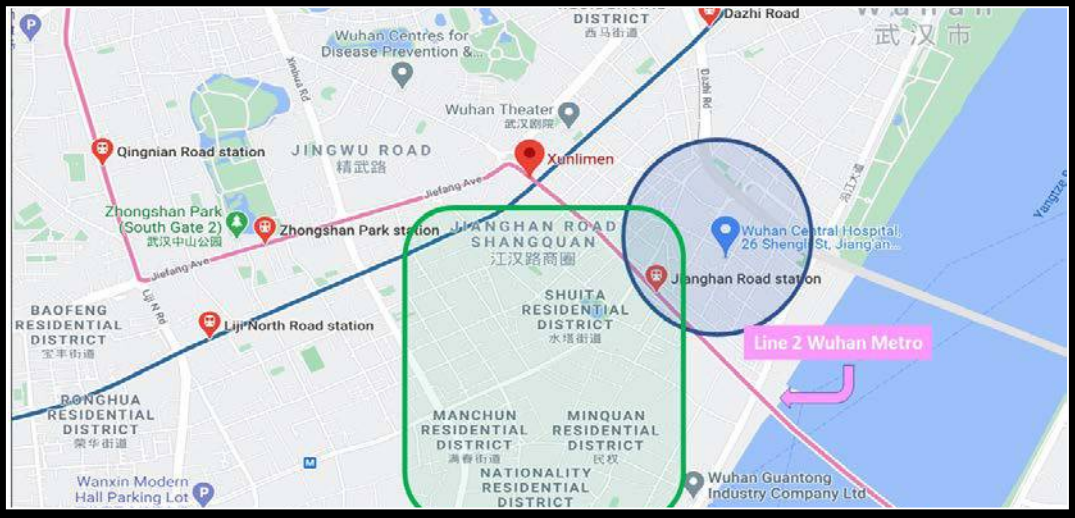
PLA Hospital is part of the Joint Logistic Support Force Complex

Position in RS	Bat-SL-CoVZC45	Bat-SL-CoVZXC21	RaTG13	PLA-4	PLA-3	PLA-2	Hu-1 Ref seq	PLA-1	GISAID #1
5' UTR	1-5 missing	1-5 missing	1-15 missing	1-16 missing	1-20 missing	1-36 missing	Intact	1-25 missing	Intact
3778	A	A	A	A	A	A	A	A	G
6968	T	T	C	C	C	C	C	A	C
8782	T	T	T	T	C	C	C	C	C
8987	T	T	T	T	T	T	T	T	A
11764	T	T	T	T	T	NA - Note 1	T	A	T
28144	C	C	C	C	T	T	T	T	T
3' UTR	last 4 poly-A missing	last 4 poly-A missing	last 13 poly-A missing	last 15 poly-A missing	last 15 poly-A missing	NA - Note 1	Intact	last 12 poly-A missing	last 4 poly-A missing
Genome length	29802	29732	29855	29872	29868	NA - Note 1	29903	29866	29899
Clade A SNPs	Clade B SNPs	Non-RaTG13 DNPs							
Note 1 - GISAID record: "Long stretches of NNNs (34.45% of overall sequence). Gap of 13 nucleotide(s) found at refpos 26171 (FRAMESHIFT). Gap of 13 nucleotides when compared to the reference sequence. 0.40% Unique Mutations."									

The PLA patient cluster

- PLA-4 is genetically the closest human infection to the three closest bat viruses
- The four PLA patients have the close sequence pattern usually seen only in family transmissions

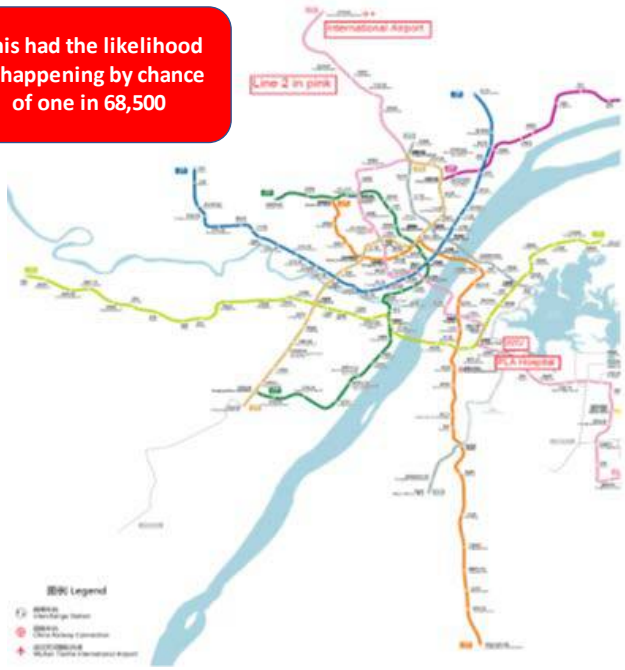
The Wuhan Metro System



Methodology: Metro station line for intersection of hospital catchment zone and residential living district zone was identified

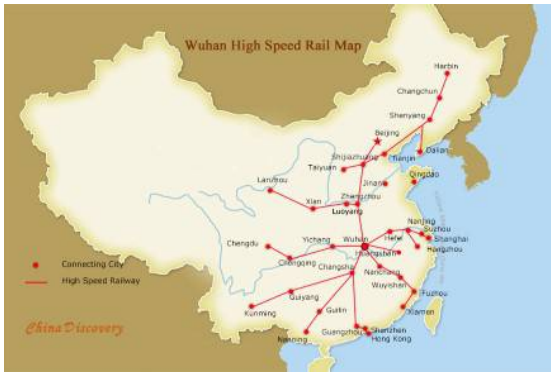
All COVID patients from Dec 1 to early Jan were admitted to hospitals on Metro Line 2

This had the likelihood of happening by chance of one in 68,500



Feature	Relationship to Pandemic
Line 2 carried 1 MM passengers a day before COVID	Assuming 2 trips/d for commuters, about 5% of the Wuhan population uses this Line, making it an efficient transmission route for all of Wuhan as well as Hubei Provence. A single patient can leave a droplet/aerosol cloud for hours to infect others.
Line 2 shares stations with every other Metro Line	Permits human-to-human spread to every part of Wuhan at the stations shared with Line 2
Line 2, Hankou Railway Station	Connects Wuhan to all of China by high speed rail
Line 2, Tianhe International Airport	International destinations: New York City, San Francisco, London, Tokyo, Rome, Istanbul, Dubai, Paris, Sydney, Bali, Bangkok, Moscow, Osaka, Seoul, and Singapore.

The Line 2 COVID Conduit



The Line 2 COVID Conduit

- Line 2 Hankou High Speed Railway Station to all of China
- Line 2 Tianhe International Airport Station to the world

The Hunan Seafood Market, Wuhan Institute of Virology, and the Wuhan CDC, all locations suggested to be the possible source of SARS-CoV-2 in Wuhan, are also all serviced by Line 2 of the Metro system, suggesting this public transit line should become the focus for further investigations into the origin of this pandemic.

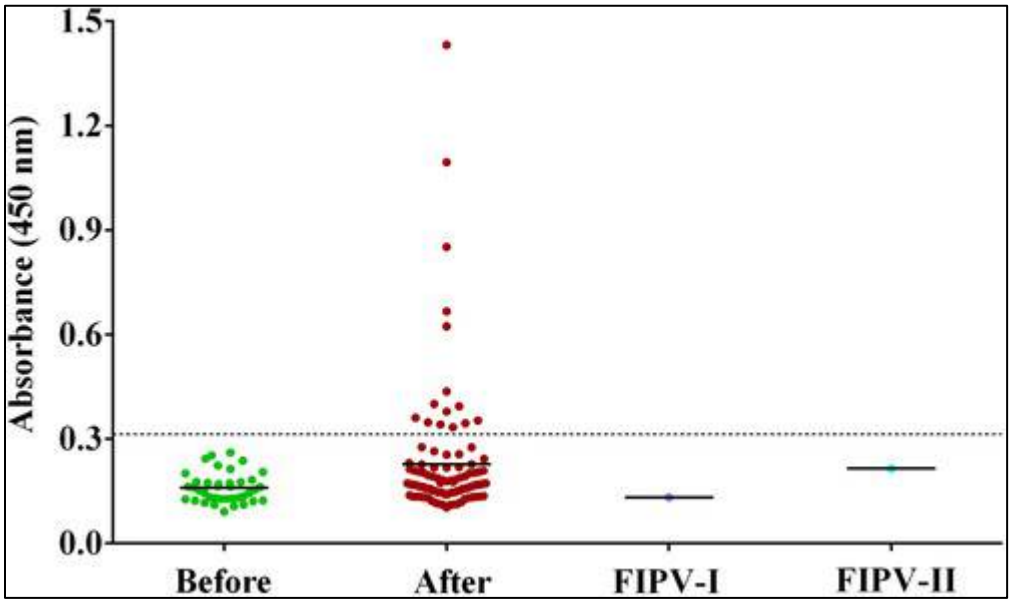
Given that the Hunan Seafood Market has been removed as a source for the origin of CoV-2, this evidence will contribute a 51%/49% contribution in favor of laboratory compared to zoonotic origin. There will be no Subjective Discount Factor adjustment.

The results from the calculations are shown below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.002	0.998
This is the outcome favors LO over ZO at 51% versus 49%		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.998 = 1.039
Normalize this step of analysis	0.002/(0.002 + 1.039) = 0.002	1.039/(0.002 + 1.039) = 0.998

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: SARS-CoV-2 infection, based on antibody seroconversion, was not found in 39 archived specimens taken from cats (1/3 feral) between March and May 2019[115]



Based on these results, the prevalence of SARS-CoV-2 in domestic and feral cats prior to January 2020 is less than 8% with a 90% confidence interval.

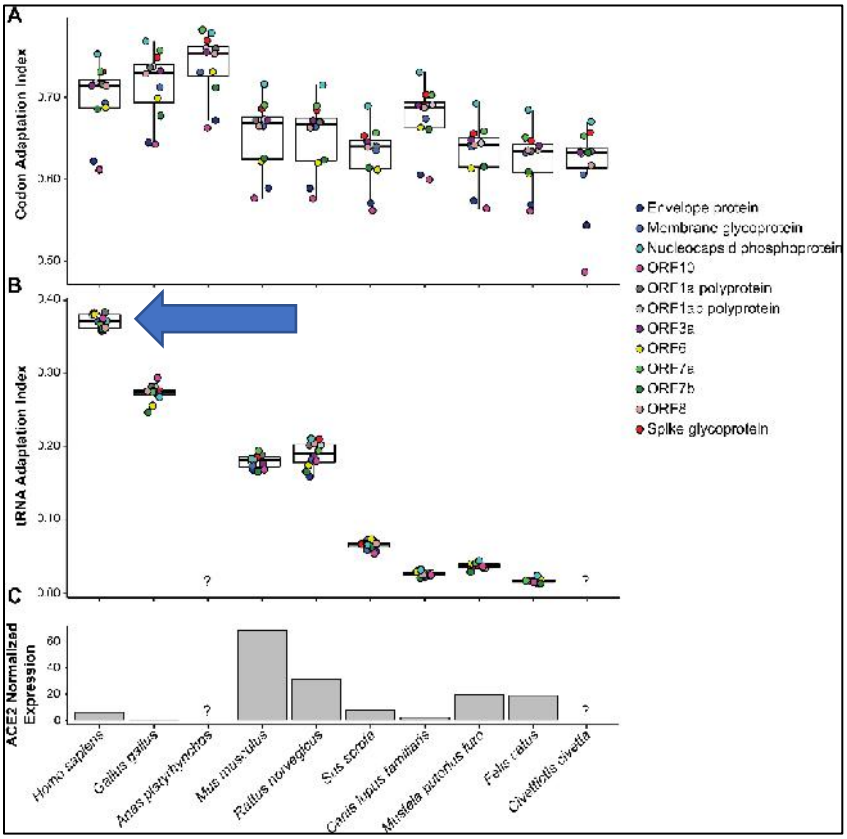
This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[115] https://www.tandfonline.com/doi/full/10.1080/22221751.2020.1817796

Evidence: The extraordinary pre-adaption of SARS-CoV-2 for human cells is demonstrated by a paper looking at a tRNA adaption index.[116]

"The proteome of SARS-CoV-2 is mainly composed of the replicase polyprotein (ORF1ab) and of structural proteins: the spike glycoprotein, the membrane and envelope proteins, and the nucleoprotein [41]. Based on the genomic codon usage of each of the possible host species, we compute the codon adaptation index (CAI) and the tRNA adaptation index (tAI) to estimate the translational efficiency of SARS-CoV-2 proteins in each host (Fig 3A and 3B and S2 Table). Humans are among the top three species whose CAIs are mostly over 0.70, together with ducks and chickens. In terms of the tAI, humans show the highest translational adaptation among all others, followed by chickens, and, to some extent, mice and rats. On the other hand, cats, ferrets, pigs, and dogs are less translationally adapted than humans both by CAI and tAI."



As shown in panel B above, the tRNA Adaption Index is highest, by far, for humans (blue arrow) followed by the red junglefowl. This is additional evidence of the extraordinary adaption of SARS-CoV-2 to humans from the very beginning. This also is the first evidence of a reasonable intermediate host but based only on these *in silico* data.

This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[116] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008450#pcbi.1008450.s004

Evidence: **Evidence of Lax procedures and disregard of laboratory safety protocols and regulations in China, including the Wuhan Institute of Virology**

A collection[117] from the Chinese Q&A website, https://www.zhihu.com/, of first-hand documentation of laboratory safety breaches and incidents within a large number of laboratories with diverse research subjects and purposes in the People's Republic of China (PRC) is provided. The laboratories involved include Chemistry labs, Biolabs, Computer labs as well as Physics and Engineering labs.

From this first-hand documentation, we obtained evidence of relaxed safety regulations and frequent breaches of such regulations, with reasons ranging from poor training/education on lab safety and chronic ignorance of safety rules, to intentional breaches of protocols for purposes other than the research projects of the lab(s) of which the breach was documented in.

Such breaches often resulted in safety accidents ranging from physical injury, chemical burns, chemical leaks, and damage to property, to lab-acquired infection and escape of in-lab pathogens. With consequences ranging from personal-level to institution-wide impacts.

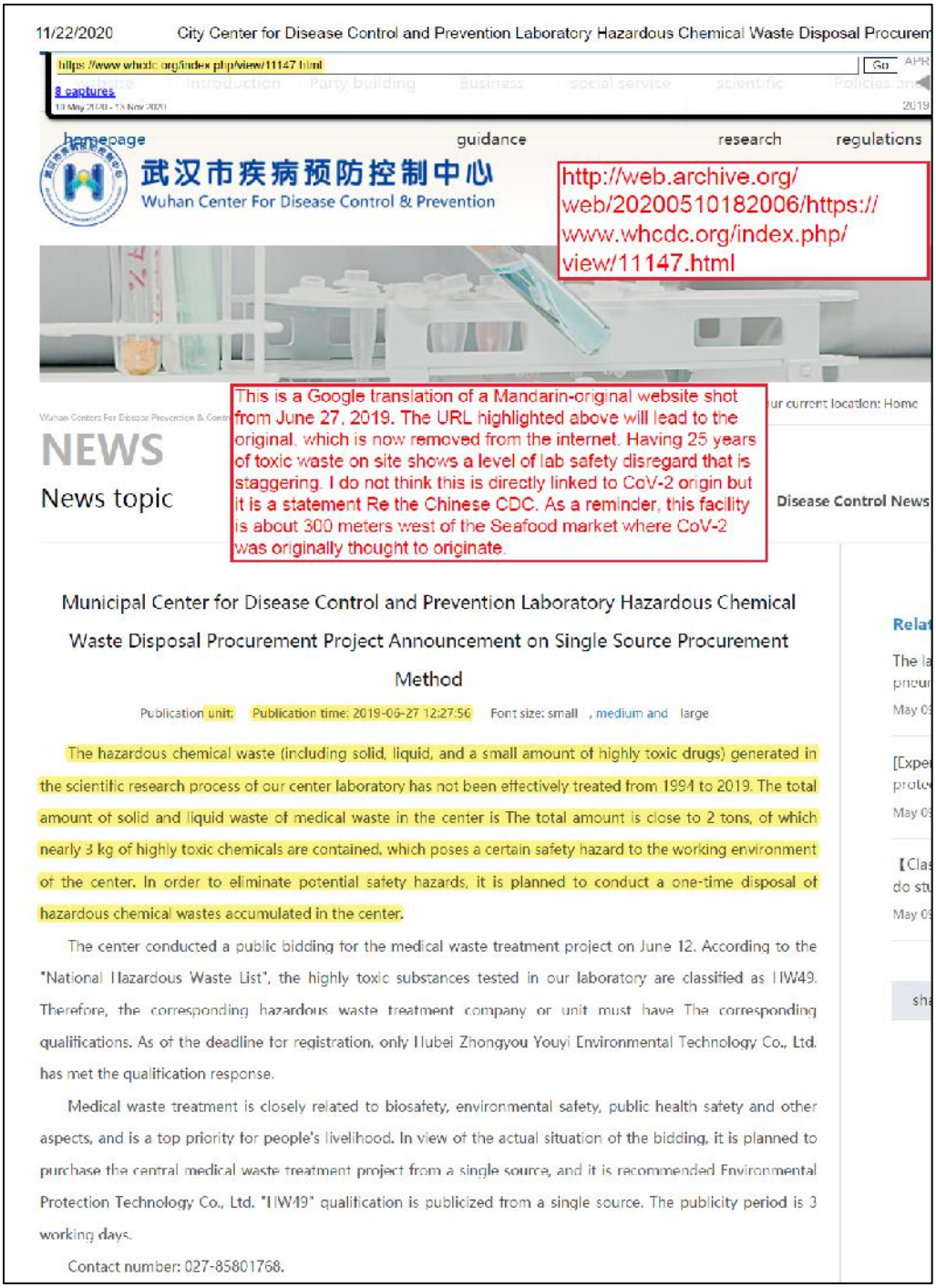
Here is the reference to the State Department cables concerning safety concerns at the WIV.[118]

The following document shows that in June 2019, the Chinese CDC was soliciting for the removal of 25-years-worth of solid and liquid medical waste. The total weight is close to two tons including three kg of highly toxic waste.

This is a Google translation of a Mandarin-original website shot from June 27, 2019. The URL highlighted above will lead to the original, which now has been removed from the internet. Having 25 years of toxic waste on site shows a staggering level of disregard for lab safety.

I do not think this is directly linked to CoV-2 origin, but it is a statement about the Chinese CDC. As a reminder, this facility is about 300 meters west of the Seafood market where CoV-2 was first thought to have originated.

[117] https://zenodo.org/record/4307879#.X-yUo9gzbOh
[118] https://foia.state.gov/Search/Results.aspx?caseNumber=F-2020-05255

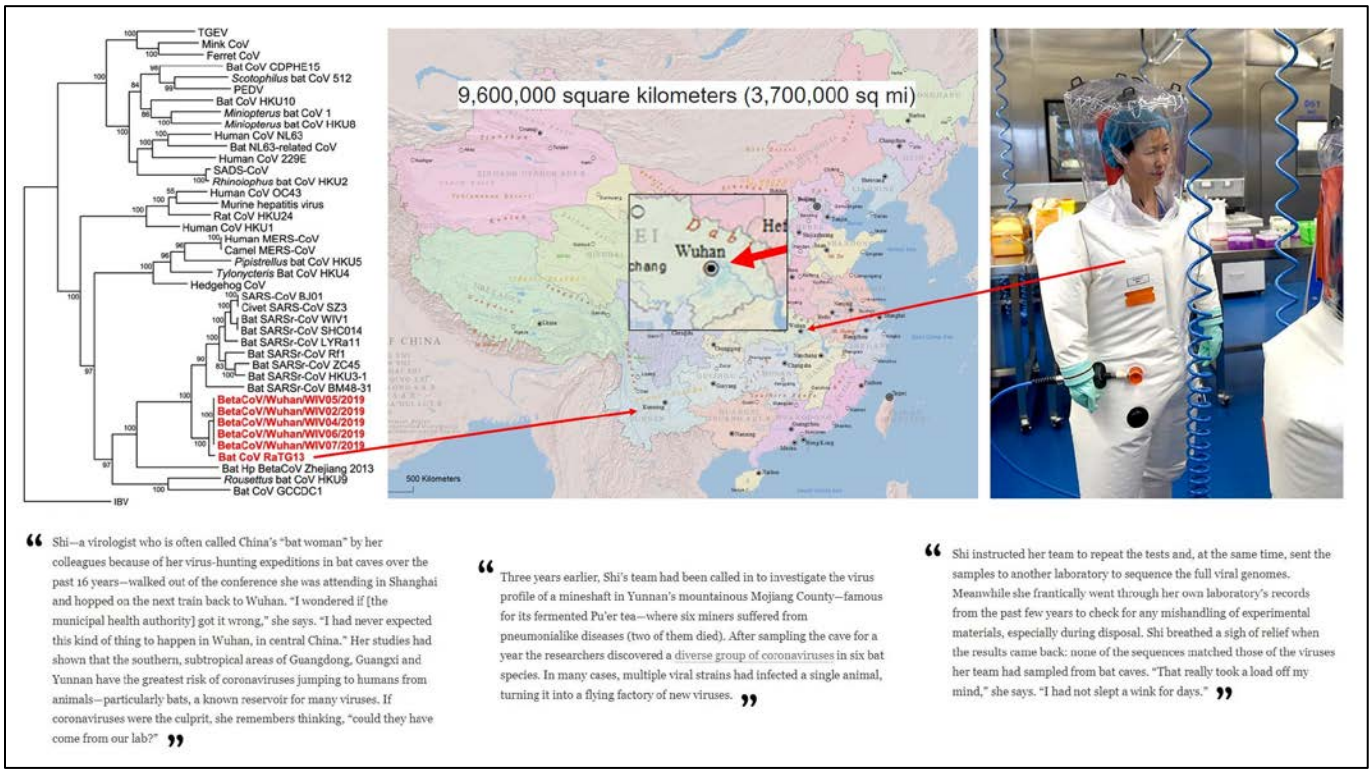


This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: The careful words of Dr. Shi do NOT say she did not have SARS-CoV-2 at the WIV.

This Figure contains quotes from an article about Dr. Shi and her reaction to the beginning of the COVID-19 pandemic.



Notice in the last frame Dr. Shi says two strange sentences:

Sentence 1: "…she frantically went through her own laboratory's records from the past few years to check for any mishandling of experimental materials, **especially during disposal**."

Why did she mention disposal? If you don't know what you are looking for this, "especially during disposal," is a bit of an odd qualifier. Other evidence from Wuhan suggests that, in fact, disposal may have been a likely source of the accidental lab release.

Sentence 2: "She breathed a sigh of relief when the results came back: none of the sequences matched those of the viruses her team had sampled from bat caves."

If Dr. Shi had created SARS-CoV-2 as a chimera, perhaps starting with one of those cave viruses, of course you would no longer have a sequence match. This is a probably truthful statement that leaves open the question of lab creation.

This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

Evidence: The Good, the Bad, and the Ugly: a review of SARS Lab Escapes[119]

In 2003–04, in the wake of the SARS epidemics, there were multiple cases of laboratory acquired infection (LAI) with SARS within just a few months: first in a P3 in Singapore, then in a military P4 in Taipei and last a protracted case in a P3 in Beijing. The 'WHO SARS Risk Assessment and Preparedness Framework' has a good summary of these lab accidents:

Since July 2003, there have been four occasions when SARS has reappeared. Three of these incidents [note: Singapore, Taipei and Beijing] were attributed to breaches in laboratory biosafety and resulted in one or more cases of SARS. The most recent laboratory incident [note: in Beijing] resulted in 9 cases, 7 of which were associated with one chain of transmission and with hospital spread. Two additional cases at the same laboratory with a history of illness compatible with SARS in February 2004 were detected as part of a survey of contacts at the facility.[i.1]

This article reviews some of these cases and discusses briefly some of the insights that were gained from these at the time.

Another article along the same lines is, "10 incidents discovered at the nation's biolabs"[120] This included Dr. Baric's laboratory in which "(b)etween April 2013 and September 2014, eight individual mouse escapes were reported at the University of North Carolina-Chapel Hill. Several of the mice were infected with either SARS or the H1N1 flu virus."

Dozens of holes in BSL-4 'spacesuits'

As a key protection against the world's most deadly pathogens, including the Ebola virus, scientists in the BSL-4 labs at the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) at Fort Detrick in Maryland wear pressurized, full-body spacesuit-like gear and breathe purified air. Yet those suits ruptured or developed holes in at least 37 incidents during a 20-month period in 2013 and 2014, according to lab incident reports obtained by USA TODAY under the federal Freedom of Information Act.

This will contribute a 51%/49% contribution in favor of laboratory compared to zoonotic origin. There will be no confidence adjustment. The results from the calculations are shown below.

Evidence or process	Zoonotic Origin (ZO)	Laboratory Origin (LO)
Starting likelihood	0.011	0.989
The history of SARS laboratory accidents is consistent with the laboratory origin hypothesis		0.51
Impact of this evidence		Increases the likelihood of LO by 51/49 = 1.041
Impact of evidence calculation		1.041 x 0.989 = 1.030
Normalize this step of analysis	0.011/(0.011 + 1.030) = 0.011	1.030/(0.011 + 1.030) = 0.989

Adjusted likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[119] https://gillesdemaneuf.medium.com/the-good-the-bad-and-the-ugly-a-review-of-sars-lab-escapes-898d203d175d
[120] https://www.usatoday.com/story/news/2015/05/29/some-recent-us-lab-incidents/25258237/

Evidence: Drs. Shi and Daszak use Wuhan residents as negative controls for zoonotic coronavirus seroconversion[121]

"As a control, we collected 240 serum samples from random blood donors in **Wuhan >1000 km away from Jinning & where inhabitants have a much lower likelihood of contact with bats due to its urban setting**" [emphasis added]. As expected, 0/240 samples from the patients from Wuhan had a positive serological evidence of prior coronavirus infection.

“The 2.7% seropositivity for the high-risk group of residents living in close proximity to bat colonies suggests that spillover is a **relatively rare event**, however this depends on how long antibodies persist in people, since other individuals may have been exposed and antibodies waned.”

In this paper from 2018, Drs. Shi and Daszak conclude that bat-to-human transfer is relatively rare for high-risk people living in close proximity to bat colonies and much less likely in Wuhan, a conclusion that does not support a hypothesis of bat-to-human transmission.

This will not be used to adjust the likelihoods.

Current likelihood: Zoonotic origin (0.2%), laboratory origin (99.8%).

[121] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6178078/

Evidence. The Bat Coronavirus RaTG13 has the Unique Genome Sequences Necessary to be the Precursor of SARS-CoV-2 Using the 'No See 'Em' Synthetic Biology Technology. ***The probability that RaTG13 acquired these 'No See 'Em' synthetic biology assembly sequences in nature is one in a billion.***

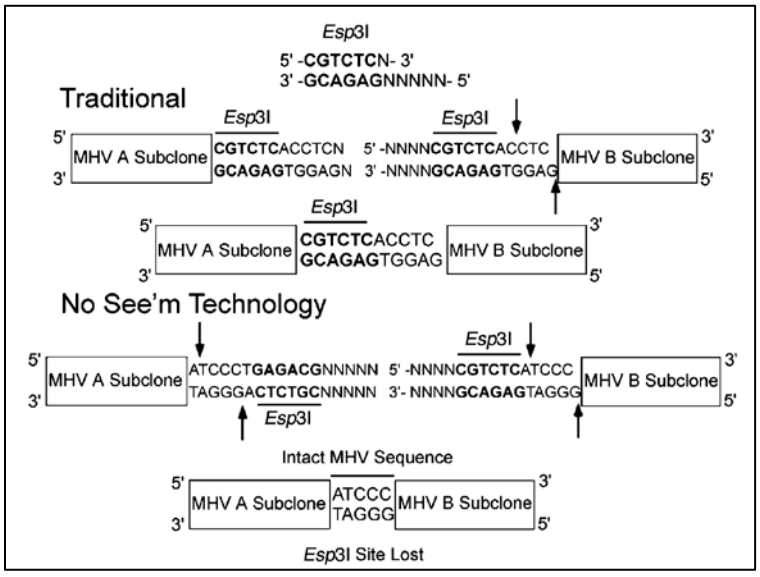
Summary.

- Synthetic biology techniques, like the engineered "No See 'Em'[122] restriction enzyme-enabled insertion method,[123] have been developed that, by design, extinguish the fingerprints of the insertion when only looking at the final genome.

- The use of these techniques is revealed however, if the precursor-product genome pair of such an insertion is available for inspection.

- **Hypothesis: the unique features of the SARS-CoV-2 Spike Protein, the receptor binding domain ACE2 contact amino acid residue region and the polybasic (furin) cleavage site, are the product of a genome insertion sequence into RaTG13 using engineered Esp3I restriction enzyme sites, the so-called, 'No See 'Em,' technology.**

- An example of the 'No See'm' Technology is shown below, taken from Baric and Sim.[1] By placing the restriction sites symmetrically on both strands of the cDNA, the resulting insertion no longer contains the identifying restriction site nts.



- According to Baric and Sims[1] "the type IIS restriction enzyme, Esp3I, recognizes an asymmetric sequence and makes a staggered cut 1 and 5 nucleotides downstream of the recognition sequence, leaving 256, mostly asymmetrical, 4-nucleotide overhangs

[122] Variably spelled 'No See 'Em,' 'No See 'um,' and 'No See'm.'
[123] https://www.researchgate.net/publication/8119695_Development_of_mouse_hepatitis_virus_and_SARS-CoV_infectious_cDNA_constructs

(GCTCTCN#NNNN). As identical Esp3I sites are generated every ~1,000,000 base pairs or so in a random DNA sequence, most restricted fragments usually do not self-assemble."

- Examination of RaTG13 identified two Esp3I cleavage sites in the Spike Protein gene, at nts 1366 and 2941 (positions 22,910 and 24,485 in the entire genome).

- As expected from the above rarity of such sites in an approximately 3800 nt gene, SARS-CoV-2 has no Esp3I sites in its SP gene. Neither do twelve other coronaviruses, including SARS-CoV-1, MERS, and other related human or bat coronaviruses.

- From all of the species other than bat RaTG13 gene source, the frequency of Esp3I sites **at any location** is 2 in 54,131 nucleotides or 0.000036947. If we assume the possibility of the occurrence of such a site at a given nucleotide is independent of any other nucleotide, then it is possible to use a binomial distribution calculation to determine the probability of 2 Esp3I sites in 3809 nucleotides for the bat RaTG13 gene. This calculation yields a probability of at least 2 sites anywhere in the Spike Protein gene of 0.009 or about one in a hundred. The probability of exactly 2 sites is 0.0086.[124]

- The 5' restriction site in RaTG13 begins at aa residue 455L, identified by Andersen et al, Nature, 2020. as the start of the "receptor-binding domain ACE2 contact residues." The downstream amino acids from this site are critical for why RaTG13 has such poor affinity for human ACE2 and the substitutions in CoV-2 are precisely why CoV-2 has such high affinity for human ACE2, why CoV-2 seems so 'preadapted' to human infections, etc. So this is the most important part of CoV-2 in explaining its ACE2 binding and infectivity. Further downstream is arguably the second most important site, the polybasic (furin) cleavage site.[125] Polybasic cleavage sites have not been observed in related 'lineage B' betacoronaviruses,' according to Andersen et al, Nature, 2020. and so there has been much speculation about how this site was acquired.

- The 3' restriction site in RaTG13 is at residue 980L. There is no protein-based rationale for this position.

- Comparing the nt sequences between RaTG13 and CoV-2, at the 5' restriction site, they are two codons in which only 2 of 6 nt bases are shared but, despite this low nt sequence homology, they are in fact synonymous base substitutions.

- Comparing the nt sequence between RaTG13 and CoV-2 at the 3' restriction site, this site has 5 of 6 identical nts with a single synonymous change in CoV-2 which destroys the restriction site. This is the only such five nt site in the RaTG13 spike protein gene and so

[124] Statistical analysis provided by Dr. Martin Lee, PhD, Adjunct Professor of Statistics, UCLA Fielding School of Public Health, UCLA, Los Angeles, CA.
[125] https://www.biorxiv.org/content/10.1101/2020.08.26.268854v1

is the easiest site in which a one nt substitution can create or destroy an Esp3I restriction site.

- The probability of having the restriction sites at **exactly these locations** can also be calculated.[2] Since there are 3809 nucleotides in the RaTG13 genome then, 3807 would not have a restriction site with probability (1-0.000036947), which was determined from the frequency of these restriction sites in other species. The other two sites would have this restriction site with probability 0.000036947. So the overall probability of this configuration has a probability of: $(1\text{-}0.00036947)^{3807} \times (0.000036947)^2 = 3.343 \times 10^{-10}$. This is a frequency of these site at their exact location being here from a natural process of approximately one in a billion.

- Dr. Zhengli-Li Shi, of the Wuhan Institute of Virology, collected the bat virus RaTG13 in 2013 and sequenced it between 2014 and 2018. In 2015, Dr. Shi and colleagues have also used the 'No See 'Em' technology' with a similar restriction enzyme, BgII, in the SARS-CoV reverse genetics system to generate chimeric coronaviruses. In that paper, they inserted a spike protein gene from a bat coronavirus into a mouse-adapted coronavirus, with a 'gain-of-function' phenotypic change.[126]

- **In conclusion:**

- **The bat coronavirus RaTG13 has two rare, Esp3I restriction sites strategically located to permit insertion of a genetic sequence that codes for the unique features of the SARS-CoV-2 Spike Protein, its receptor binding contact amino acids and its polybasic (furin) cleavage site, using the 'No See 'Em' synthetic biology techniques.**

- **This specific synthetic biology laboratory technique has been successfully performed previously by Wuhan Institute of Virology scientists to increase coronavirus infectivity.**

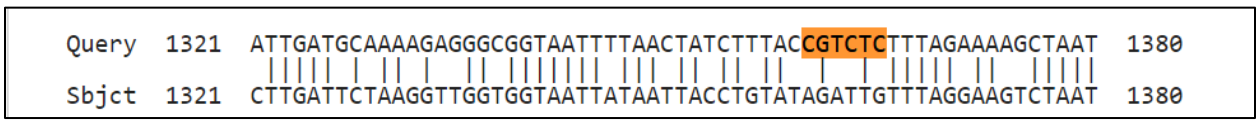
- **The probability these two sites are present and in their exact location in RaTG13 by an act of nature is one in a billion.**

[126] https://www.nature.com/articles/nm.3985

Text-Table. A record of the EspI restriction enzyme sites in the Spike Protein (SP) genes of fifteen coronaviruses, including RaTG13 and SARS-CoV-2. RaTG13 is unique in having two such sites, with SARS-CoV-2 and eleven other coronaviruses having no such site in the SP gene. The restriction sites were identified with the RestrictionMapper site algorithm: http://www.restrictionmapper.org/ .

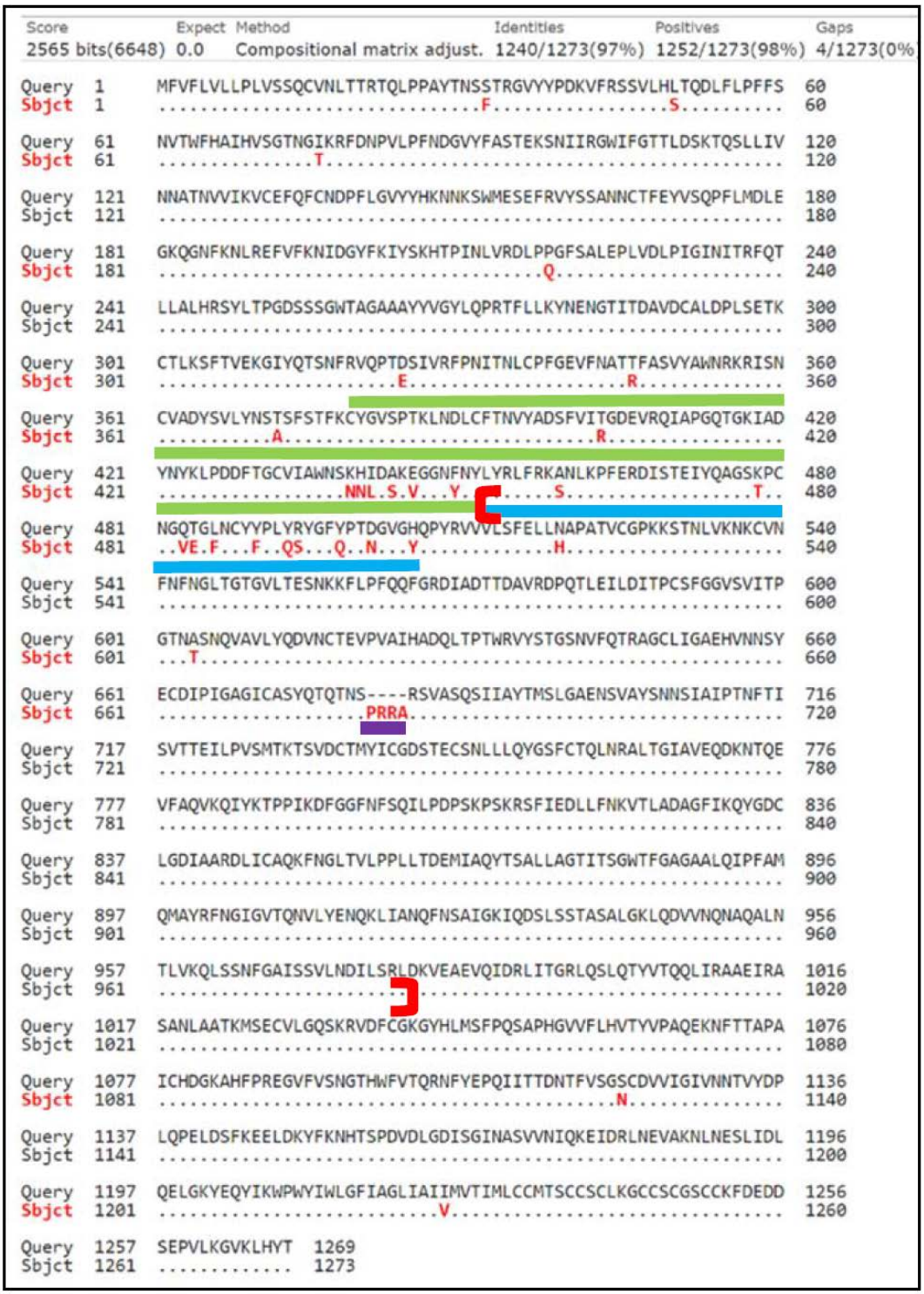
Species	Spike Protein (SP) Gene Source	Nt Size of SP Gene	Esp3I Site Location in Spike Protein Gene	Reference
Bat	Bat Coronavirus RaTG13 from WIV	3809	1366, 2941 (22910, 24485 in genome)	
Human	SARS-CoV-2 Reference Sequence	3821	None	
Bat	Rhinolophus affinis coronavirus isolate LYRa11	3779	None	Daszak and Shi paper
Bat	Bat SARS coronavirus HKU3-1	3728	None	Daszak and Shi paper
Bat	SARS-like coronavirus isolate bat-SL-CoVZC45	3740	None	Third Military University publication
Bat	SARS-like coronavirus bat-SL-CoVZXC21	3737	None	Third Military University publication
Bat	hCoV-19/bat/Yunnan/RmYN02/2019	3873	None	Wild bat coronavirus with apparent furin-like insert
Bovine	Bovine coronavirus strain Quebec	4091	None	
Human	Human coronavirus HKU1 strain	4070	3208	
Human	MERS Reference Sequence	4061	None	
Human	Human coronavirus OC43 strain	4079	None	
Human	Human coronavirus 229E strain	3512	None	
Human	Human Coronavirus NL63 Reference Sequence	4070	None	
Human	SARS 2003 coronavirus ZJ0301	3767	None	
Pangolin	Pangolin coronavirus isolate PCoV_GX-P4L	3803	3351	
Human	SARS-CoV-1 Urbani	3767	None	

Figure. A comparison of the RaTG13 Spike Protein gene (Query) and the SARS-CoV-2 Reference Sequence (Sbjct) showing the only two Esp3I restriction enzyme cleavage site, both present in RaTG13 but absent in SARS-CoV-2. The restriction sites were identified with the RestrictionMapper site: http://www.restrictionmapper.org/ .The 5' cleavage site is strategically located at the beginning of the receptor binding domain ACE2 contact residues. Despite four of six nt are different these are synonymous changes.



The 3' cleavage site is the only downstream -CGTCTN- sequence found in the CoV-2 Spike Protein, making it unique.

Figure. Comparison of Spike Protein amino acid sequence between RaTG13 (Query) and SARS-CoV-2 (Sbjct). Amino acid substitutions in CoV-2 are shown in red, single letter abbreviation. Green band; receptor binding domain. Blue band; receptor binding domain ACE2 contact residues (Andersen et al, Nature, 2020.). Purple band; polybasic (furin) cleavage site. Red brackets; Esp3I cleavage sites in RaTG13.



```
 Score          Expect  Method                    Identities          Positives           Gaps
 2565 bits(6648)  0.0   Compositional matrix adjust.  1240/1273(97%)  1252/1273(98%)  4/1273(0%

Query  1     MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSSTRGVYYPDKVFRSSVLHLTQDLFLPFFS  60
Sbjct  1     ..............................F.................S.........  60

Query  61    NVTWFHAIHVSGTNGIKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV  120
Sbjct  61    ...............T.............................................  120

Query  121   NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE  180
Sbjct  121   ............................................................  180

Query  181   GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPPGFSALEPLVDLPIGINITRFQT  240
Sbjct  181   ......................................Q.....................  240

Query  241   LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK  300
Sbjct  241   ............................................................  300

Query  301   CTLKSFTVEKGIYQTSNFRVQPTDSIVRFPNITNLCPFGEVFNATTFASVYAWNRKRISN  360
Sbjct  301   ......................E.....................R..............  360

Query  361   CVADYSVLYNSTSFSTFKCYGVSPTKLNDLCFTNVYADSFVITGDEVRQIAPGQTGKIAD  420
Sbjct  361   ...........A................................R..............  420

Query  421   YNYKLPDDFTGCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPC  480
Sbjct  421   .................NNL.S.V...Y.....S..................T..  480

Query  481   NGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATVCGPKKSTNLVKNKCVN  540
Sbjct  481   ..VE.F...F..QS...Q..N...Y.............H.....................  540

Query  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600
Sbjct  541   ............................................................  600

Query  601   GTNASNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
Sbjct  601   ...T........................................................  660

Query  661   ECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716
Sbjct  661   .....................PRRA.....................................  720

Query  717   SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  776
Sbjct  721   ............................................................  780

Query  777   VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  836
Sbjct  781   ............................................................  840

Query  837   LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  896
Sbjct  841   ............................................................  900

Query  897   QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  956
Sbjct  901   ............................................................  960

Query  957   TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1016
Sbjct  961   ..................................................................  1020

Query  1017  SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1076
Sbjct  1021  ............................................................  1080

Query  1077  ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDP  1136
Sbjct  1081  ..........................................N.................  1140

Query  1137  LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL  1196
Sbjct  1141  ............................................................  1200

Query  1197  QELGKYEQYIKWPWYIWLGFIAGLIAIIMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD  1256
Sbjct  1201  ............................V...............................  1260

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

Because it has not been established that RaTG13 was the precursor of CoV-2 this evidence statement will not be used at this time to adjust the likelihoods of the origin. If additional information is obtained at a later date this may be revisited.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (0.2%) and laboratory origin (99.8%)

Evidence. Location, location, location: Based on the distance between known SARS-CoV-1 laboratory-acquired infections and the hospital of admission of the infected personnel, the WIV is within the expected hospital catchment for a CoV-2 LAI

Hypothesis. Laboratory-acquired infections (LAI) have the property that the hospital of admission of the personnel from the laboratory with the acquired infection is close to the laboratory, specifically they are within 24.64 km (95% Confidence Interval) from the laboratory.

Prior data from SARS-CoV-1. There were four LAIs of SARS-CoV-1 that can be used to determine the distance between the laboratory where the infection occurred and the hospital of first admission. The data are here:

SARS-CoV-1 Laboratory Acquired Infection (LAI)	Hospital of admission	Distance (Google Maps)	
In September 2003, a 27-year-old student from the National University of Singapore (NUS) was infected with the SARS virus due to improper experimental procedures	Singapore General Hospital (SGH)	6.3 km	
Baiji Mountain, Sanxia, Taiwan	Taiwan Hoping Hospital, Taipei, Taiwan	27.8 km	
№100 Yingxin Street, Xicheng District, Bejing	Union Hospital, Beiijing, China	7.3 km	
№100 Yingxin Street, Xicheng District, Bejing	Friendship Hospital, Beijing, China	17.6 km	
		mean = 14.75	
		SD = 10.1	
		95% Confidence Interval	14.75 ±9.887

Based on these four cases, the 95% upper confidence limit for the distance from LAI patients to the hospitals of admission is 24.6 km of the laboratory where the infection was acquired.

SARS-CoV-2. Although it is not clear which hospital the first patient was admitted to the following Text-Table contains all likely candidates.

SARS-CoV-2 Potential LAI Source	Hospital of admission	Distance (Google Maps)	Probability of being closer than the average results for SARS-CoV-1	Probability of being farther than the average results for SARS-CoV-1
Wuhan Institute of Virology, Wuhan, China	PLA Hospital, NO. 627 Wuluo Road, Wuchang District, Wuhan, China	4.8 km	0.094	0.906
Wuhan Institute of Virology, Wuhan, China	Wuhan Central Hospital, Wuhan, China	9.1 km	0.338	0.662
Wuhan Institute of Virology, Wuhan, China	Zhongnan Hospital, Wuhan, China	2.8 km	0.019	0.981
Wuhan Institute of Virology, Wuhan, China	Tongji Hospital, Wuhan, China	5.1 km	0.109	0.891
Wuhan Institute of Virology, Wuhan, China	Hubei Maternity and Child Health Care Hospital, Wuhan, China	4.4 km	0.075	0.925
Hypothesis: Given the distance from the SARS-CoV-1 laboratory where an LAI occurred to the hospital of admission for the lab workers who became infected, what is the probability that CoV-2 is also an LAI, given the distance from the hospitals where the first patients were seen to the WIV, the hypothesized source.			Probability calculations based on the use of a log-normal distribution for distances	Probability calculations based on the use of a log-normal distribution for distances

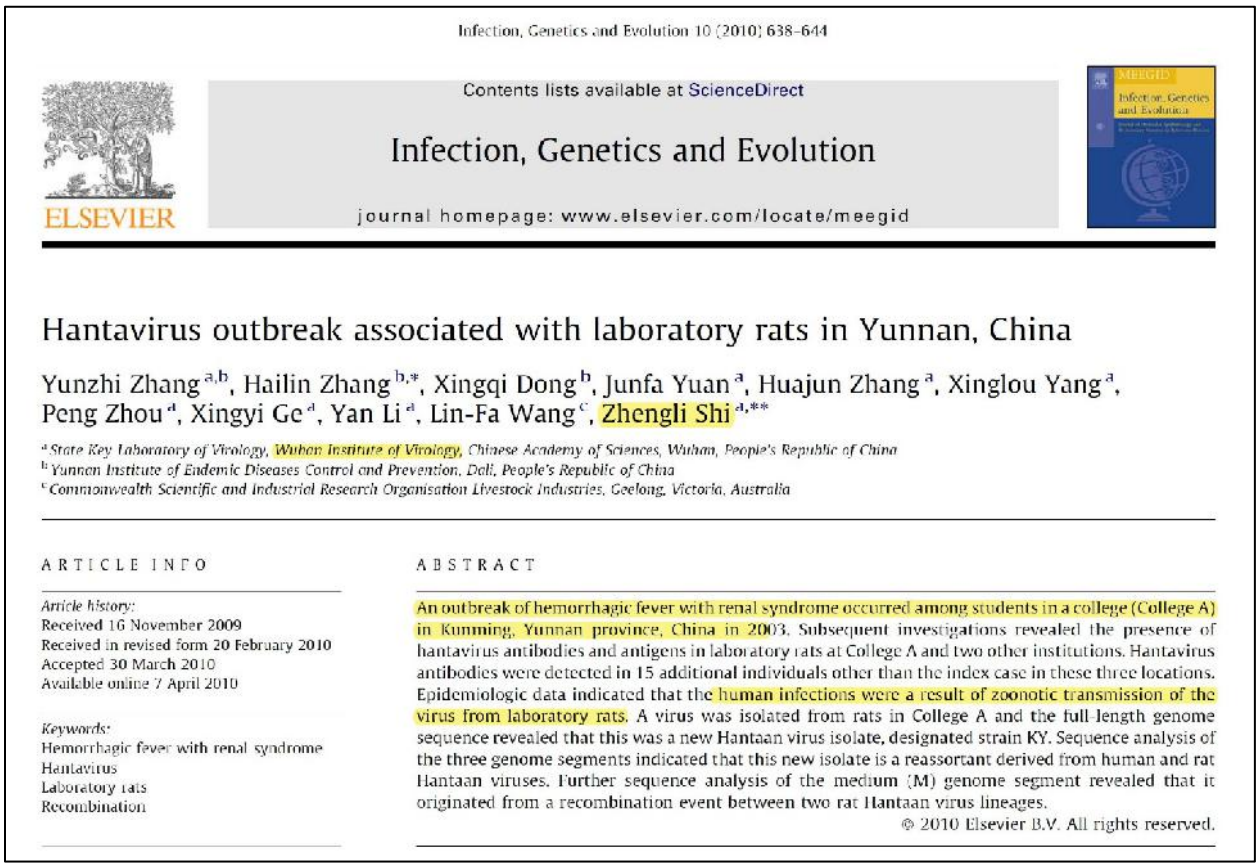
Based on the data for actual LAI for SARS-CoV-1 the distance between the WIV and the hospitals of admission for CoV-2 is consistent with the WIV being the origin for the LAI. There is no evidence the putative LAI for CoV-2 is any different than the known LAIs for CoV-1.

This evidence is not independent of other evidence that is based on location and so it cannot be used independently in the Bayesian analysis. It is included here for completeness.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (0.2%) and laboratory origin (99.8%)

Evidence. Dr. Shi successfully identifies a laboratory-acquired infection outbreak from Hanta virus in laboratory rodents.



Infection, Genetics and Evolution 10 (2010) 638–644

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Hantavirus outbreak associated with laboratory rats in Yunnan, China

Yunzhi Zhang [a,b], Hailin Zhang [b,*], Xingqi Dong [b], Junfa Yuan [a], Huajun Zhang [a], Xinglou Yang [a], Peng Zhou [a], Xingyi Ge [a], Yan Li [a], Lin-Fa Wang [c], Zhengli Shi [a,**]

[a] State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, People's Republic of China
[b] Yunnan Institute of Endemic Diseases Control and Prevention, Dali, People's Republic of China
[c] Commonwealth Scientific and Industrial Research Organisation Livestock Industries, Geelong, Victoria, Australia

ARTICLE INFO

Article history:
Received 16 November 2009
Received in revised form 20 February 2010
Accepted 30 March 2010
Available online 7 April 2010

Keywords:
Hemorrhagic fever with renal syndrome
Hantavirus
Laboratory rats
Recombination

ABSTRACT

An outbreak of hemorrhagic fever with renal syndrome occurred among students in a college (College A) in Kunming, Yunnan province, China in 2003. Subsequent investigations revealed the presence of hantavirus antibodies and antigens in laboratory rats at College A and two other institutions. Hantavirus antibodies were detected in 15 additional individuals other than the index case in these three locations. Epidemiologic data indicated that the human infections were a result of zoonotic transmission of the virus from laboratory rats. A virus was isolated from rats in College A and the full-length genome sequence revealed that this was a new Hantaan virus isolate, designated strain KY. Sequence analysis of the three genome segments indicated that this new isolate is a reassortant derived from human and rat Hantaan viruses. Further sequence analysis of the medium (M) genome segment revealed that it originated from a recombination event between two rat Hantaan virus lineages.

The significance of this evidence is that it demonstrates the methods used by Dr. Shi and the WIV to solve a laboratory-acquired infection outbreak. The methods described herein should be applied to the WIV in order to determine if CoV-2 was also a laboratory-acquired infection.

This will not be used to directly advance the Bayesian analysis.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (0.2%) and laboratory origin (99.8%)

Evidence. Bats hibernate when the temperature is below 10.5 C;[127] in Hubei province that begins in September and ends in May.

Average Hubei Temperature by Month

Month	Recommended Rate	Max Temp.	Min Temp.
Jan.	~	-17°C	-26°C
Feb.	√	-13°C	-23°C
Mar.	√√	-3°C	-14°C
Apr.	√√	7°C	-4°C
May.	√√	16°C	4°C
Jun.	√	23°C	11°C
Jul.	√	23°C	13°C
Aug.	√	21°C	11°C
Sep.	√√	15°C	4°C
Oct.	√√	6°C	-5°C
Nov.	√√	-5°C	-16°C
Dec.	√	-15°C	-23°C

Based on this evidence, they would have been hibernating at the time of the first human outbreak in the fall of 2019. Since this evidence is cumulative to the prior evidence from Dr. Shi that the bat host species for CoV-2 does not live in Hubei Province it will not be used to change the Bayesian analysis.

Likelihood from prior state is unchanged following this evidence analysis:

Zoonotic origin (0.2%) and laboratory origin (99.8%)

[127] https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7998.1971.tb01323.x