



OPINION *This piece expresses the views of its author(s), separate from those of this publication.*

Deleted COVID-19 genetic fingerprints show it's still possible to dig for lab leak evidence

By 'deeply probing' coronavirus data digitally archived outside China – from grant reports to reviews of scientific papers – more clues may be found.

Alison Young Opinion contributor

Published 4:57 p.m. ET Jun. 24, 2021 | Updated 5:32 p.m. ET Jun. 24, 2021

A respected U.S. scientist researching the evolution of the COVID-19 virus has uncovered an intriguing mystery with potentially troubling implications: Some of the virus' earliest genetic fingerprints were quietly deleted last year from an important international database at the request of Chinese scientists.

The deleted data raises questions about whether efforts have been made to “obscure” information in scientific databases that hold clues to knowing where the SARS-CoV-2 virus originated – and whether the pandemic started with a chance human encounter with an infected animal, or through a laboratory accident in Wuhan.

Amid the slow acceptance that the “lab leak” theory is worthy of legitimate investigation, this finding shows that it may still be possible to unearth new evidence about how the pandemic began even if the Chinese government and the Wuhan Institute of Virology won't open their records to independent investigators.

Now that people are finally looking, what else might be found in forgotten troves of archived emails, obscure databases, funding application materials and other records that have zipped back and forth over the internet between scientists in China and individuals and organizations in the United States and other countries?

Could an accident have caused COVID-19?: Why the Wuhan lab-leak theory shouldn't be dismissed

The discovery of the deleted virus gene sequences is a start.

“There is no plausible scientific reason for the deletion,” said Jesse Bloom, a leading evolutionary biologist at the Fred Hutchinson Cancer Research Center in Seattle, in a scientific paper posted late Tuesday on the pre-publication website bioRxiv, where scientists make findings immediately available before peer review.

“It therefore seems likely the sequences were deleted to obscure their existence,” wrote Bloom, who last month helped organize an influential letter from a group of scientists calling for both natural and laboratory theories of the pandemic’s origin to be taken seriously and investigated.

It’s a serious charge that has already drawn attention from the World Health Organization and the National Institutes of Health.

'There are also broader implications'

It all started with Bloom reading a journal article about SARS-CoV-2 sequences that had been deposited in the NIH database by the end of March 2020. But when he went looking for them in the NIH database, he couldn’t find them.

By digging into archives on a Google Cloud server, Bloom was able to track down 13 deleted sequences that he said provide a little more evidence the virus was circulating in Wuhan before an early epidemic at the Huanan Seafood Market that has been the focus of a recent joint WHO-China report on the pandemic’s origin.

The bigger finding, he said, was that they had been deleted.

“There are also broader implications,” Bloom said on Twitter. That the data was deleted “should make us skeptical that all other relevant early Wuhan sequences have been shared.” Bloom was not available for an interview.

In an email to me, WHO spokesperson Tarik Jašarević said, “We are aware of this report and, as we repeatedly asked, we hope that all data on early cases will be made available.”

The data was deleted from a massive, international public database maintained by the National Institutes of Health that archives millions of records on gene sequences collaboratively shared by scientists worldwide so they can be used for further research on a wide range of viruses, including SARS-CoV-2.

After initially sharing the sequences on the database in March 2020, the researcher from a team largely based in Wuhan submitted a request to the NIH in June 2020 asking that they be

removed, the NIH told me in a statement.

“NIH can’t speculate on motive beyond the investigator’s stated intentions,” the agency said.

'I remember it very well': Dr. Fauci describes a secret 2020 meeting to talk about COVID origins

The researcher indicated that the virus sequences had been updated and were being submitted to another database, and that they wanted the data removed from the NIH database to avoid confusion between the versions.

Shortly after the data was removed from the NIH database, Bloom discovered that the sequences were also removed from the China National GeneBank DataBase, another online repository.

NIH said the researcher didn’t specify what database would be receiving the updated sequences. Because it’s a voluntary and collaborative system, researchers are free to remove their sequences.

The deleted data was uploaded in association with two scientific papers, a preprint version posted online in March 2020 and a final article in a nanotechnology journal that was published in June 2020. Even though the sequences were deleted from the centralized database – called the Sequence Read Archive – that other scientists use to do their analyses, the papers that discuss the data remain online. But you would need to know to go looking for them.

“The practical consequence of removing the sequences from the Sequence Read Archive is that no one knew they existed prior to now, and they were not in the databases used to collect Dec-Jan sequences for the joint WHO-China report,” Bloom said in an email to me.

More clues could be found

I emailed the corresponding authors of both articles Tuesday night to ask why the sequences were deleted. But I have not received a reply. Bloom noted in his paper that he didn’t receive a reply, either.

The Chinese Embassy in Washington, while not addressing the deleted sequences, contends China has been transparent since the outbreak of COVID-19 began. “To politicize origin tracing, a matter of science, will not only make it hard to find the origin of the virus, but give

free rein to the ‘political virus’ and seriously hamper international cooperation on the pandemic,” an embassy spokesperson said in an email.

WHO Director-General Tedros Adhanom Ghebreyesus, along with some international members of a joint WHO-China team of experts examining the origin of the pandemic, have expressed frustration that Chinese government officials have refused to share raw data on early COVID-19 patients that might help determine where the virus came from.

Confusion and falsehoods: How we can fix the COVID trust gap and get people vaccinated

In recent weeks, there has been growing international support for investigating whether a laboratory accident in the fall of 2019 caused the pandemic – a line of inquiry long ignored because high-profile scientists branded it as a crackpot conspiracy theory.

President Joe Biden has asked U.S. intelligence agencies to increase their efforts to find the origin of the pandemic, whether it came from human contact with an infected animal – which is the common way viruses emerge in nature – or whether a lab accident was involved.

Bloom’s forensic work illustrates that there may be important evidence about the virus’ origin that can be discovered without on-the-ground cooperation of the Chinese government. By “deeply probing” data that is digitally archived outside China’s borders – everything from archived genomic data to grant reports to reviews of scientific papers – Bloom thinks more clues may be found.

He said the international scientific community also needs to determine whether collaborative databases that researchers rely on have been manipulated.

"It is important to examine if other trust-based systems in science conceivably may have also been used to hide data relevant to origins / early spread of #SARSCoV2," Bloom wrote on Twitter.

When I asked NIH about this, the agency said it has been analyzing the database and “found that eight SARS-COV-2 submission packages were withdrawn upon request of the submitter since the beginning of the pandemic. This included one requested by a submitter from China and the rest from submitters predominantly in the U.S.” Additional details weren’t provided about what those deletions involved.

It’s difficult to know what to make of all of this.

No signs of 'a malicious reason'

Steven Salzberg, who has served as an adviser to an international group that oversees these collaborative databases, said he can't speculate why the sequences were deleted, "although it does seem a bit unusual."

"Because the (NIH databases) are so enormous, the process for managing them has to be automated," said Salzberg, a professor of biomedical engineering at Johns Hopkins University. "It's entirely legitimate for a scientist to delete or replace data in one of those databases if he/she discovers that the data is erroneous or has other technical problems."

Demand accountability: Trump and raging pandemic helped China dodge COVID accountability.

But no corrections have been noted on the journal articles published with the deleted data, Bloom noted.

Bloom's discovery of the deleted virus sequences was dismissed and even derided by some scientists who have been outspoken in their belief that investigating the so-called lab leak theory of the pandemic is a waste of time that runs the risk further antagonizing Chinese authorities and ruining any hope of collaborative investigations into a natural origin of the virus.

In an example of how ugly some of the scientific debate has become on the topic, virologist Angela Rasmussen on Tuesday accused Bloom, without directly naming him, of making a weak case and essentially "asking systemic racism to do the heavy lifting for you to make your point." That tweet has since been deleted.

Rasmussen said on Twitter that the paper "relies on the premise that scientists from China are not trustworthy. There's no evidence those sequences were deleted for any malicious reason." Rasmussen, a research scientist based at the University of Saskatchewan who also has been affiliated with Georgetown University's Center on Global Health Science and Security, said in an email Wednesday evening that her opinion was based on Bloom's statements on Twitter, which she believes have framed the data "in the context of insinuations and omissions."

Bloom noted in his paper and on Twitter that the data deletions occurred in the context of orders by the Chinese government that unauthorized labs in the country destroy their virus samples early in the outbreak and other directives that required scientists receive central approvals before publishing articles about COVID-19. The bottom line: Individual scientists may have had no choice in the matter.

Through some dogged sleuthing, Bloom has shown the lab leak hypothesis – ruling it in or out – may not be a hopeless cold case. There still may be a lot more data out there now that investigators have finally started digging.

“After spending the last 4 months studying this closely, I am cautiously optimistic that additional relevant data are still likely to come to light,” he said.

Alison Young is an investigative reporter in Washington, D.C. During 2009-19, she was a reporter and member of USA TODAY's national investigative team. Follow her on Twitter: @alisonannyoung