

EMPOWER OVERSIGHT

Whistleblowers & Research



CORONAVIRUS SEQUENCES REMOVED FROM NIH DATABASE AT THE REQUEST OF CHINESE RESEARCHERS

Introduction

In contrast to best practices of scientific openness and collaboration, the National Institutes of Health (NIH) deleted information about coronavirus genetic sequences at the request of Chinese scientists in the midst of the COVID-19 global pandemic. Empower Oversight has been seeking answers since the summer of last year. After dodging questions from Congress and being sued under the Freedom of Information Act (FOIA), NIH finally produced documents shedding some light on the circumstances of the deletions.

On July 14, 2021, Empower Oversight filed a FOIA request with the NIH seeking transparency about controversial deletions from the Sequence Read Archive (SRA). NIH operates the database as part of its participation in the International Nucleotide Sequence Database Collaboration (INSDC) in order to “capture, organise, preserve and present nucleotide sequence data as part of the open scientific record.”¹ INSDC has noted that “The global COVID-19 crisis has brought an urgent need for the rapid open sharing of data relating to the outbreak.”²

On June 22, 2022, researcher Jesse Bloom, a virologist at the Fred Hutchinson Cancer Research Center, published a preprint that reported on the deletions of coronavirus sequences at the request of Chinese researchers.³ That preprint spurred several media reports⁴ and letters from United States Senators.⁵ NIH was nonresponsive to congressional oversight requests, as well as to Empower Oversight’s FOIA request about these sequence deletions. But, after Empower Oversight sued to enforce its request, the NIH produced 238 pages of documents related to the deletions and 17 pages of documents related to the Senate inquiries.⁶

¹ INSDC, “Statement on SARS-CoV-2 sequence data sharing during COVID-19” (emphasis added).

[https://www.insdc.org/sites/insdc.org/files/documents/INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19.pdf](https://www.insdc.org/sites/insdc.org/files/documents/INSDC%20Statement%20on%20SARS-CoV-2%20sequence%20data%20sharing%20during%20COVID-19.pdf)

² *Id.*

³ Jesse Bloom, “Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic,” *Molecular Biology and Evolution* (Jun 22, 2021).

<https://www.biorxiv.org/content/10.1101/2021.06.18.449051v1>

⁴ Amy Dockser Marcus and Drew Hinshaw, “After Covid-19 Data Is Deleted, NIH Reviews How Its Gene Archive Is Handled,” *The Wall Street Journal* (Sep 13, 2021).

<https://www.wsj.com/articles/after-covid-19-data-is-deleted-nih-reviews-how-its-gene-archive-is-handled-11631545490>

⁵ “Did NIH Improperly Delete COVID-19 Data At Request Of Chinese Researchers? Senators Want Answers,” Press Release (Sep 16, 2021).

<https://www.grassley.senate.gov/news/news-releases/did-nih-improperly-delete-covid-19-data-at-request-of-chinese-researchers-senators-want-answers>

⁶ “Empower Oversight Amends Complaint in NIH Lawsuit on Deleted Coronavirus Sequences,” Press Release (Mar 1, 2022).

<https://empower.us/empower-oversight-amends-complaint-in-nih-lawsuit-on-deleted-coronavirus-sequences/>

Empower Oversight is releasing the 238 pages of documents for the first time in conjunction with this report. Litigation is ongoing to obtain more records, but this research summarizes what can be learned from the initial set of NIH documents.

Key Findings

1. **Documents indicate that an expert advised Collins and Fauci that the deleted sequences may suggest the pandemic began outside the Huanan Seafood Wholesale Market in Wuhan.** After Bloom alerted NIH about the deleted sequences, NIH Director Francis Collins and Anthony Fauci, the Director of the National Institute of Allergy and Infectious Diseases, hosted a Sunday afternoon Zoom meeting. The invitation that Collins sent out for the meeting asks invitees to read Bloom’s preprint paper closely and provide their “advice on the interpretation and significance of” it. Professor Trevor Bedford of the Fred Hutchinson Cancer Research Center later sent the group an email stating that the deleted data seemed to support the idea that the pandemic began outside the Huanan market in Wuhan and that the matter must be analyzed properly.

2. **The NIH initially declined a Wuhan University researcher’s request to remove the sequences before agreeing to a second, related request and then offering to remove both sets of sequences.** On June 5, 2020, a Wuhan University researcher requested that NIH retract the researcher’s submission of BioProject ID PRJNA637497 because of error. The Wuhan researcher explained “I’m sorry for my wrong submitting.” BioProject ID PRJNA637497 is also referred to as Submission ID SUB7554642.

Three days later, on June 8th, the NIH declined the researcher’s request, advising that it prefers to edit or replace, as opposed to delete, sequences submitted to the SRA. On June 15, 2020, referring to a related submission, the same Wuhan University researcher advised:

Recently, I found that it’s hard to visit my submitted SRA data, and it would also be very difficult for me to update the data. I have submitted an updated version of this SRA data to another website, so I want to withdraw the old one at NCBI in order to avoid the data version issue. The Submission ID is SUB7147304.

The next day, NIH agreed to the request, and asked whether the Wuhan University researcher also wanted NIH to delete Submission ID SUB7554642, which NIH had refused to remove a week prior. The email states:

Do you want to withdraw all SRA objects in your account? here are 2 submissions SUB7554642 and SUB7147304. Also, bioprojects and biosamples whould [sic] be withdrawn as well, right?

The Wuhan University researcher responded, “Yes, I want to withdraw both 2 submissions” as well as all “The Bioprojects, Biosamples and all SRA objects.”

NIH then replied that it “had withdrawn everything.”

3. **NIH appears to have misled reporters about the policy for removing sequences.** On June 19, 2021, an NIH official from the information and engineering branch wrote in an internal email, “The only way data is removed from the SRA (per SOP) is if a submitter notifies us that the submission was in error.” However, that was not the stated grounds for the June 2020 removal of the genetic sequences identified as Submission IDs SUB7147304 and SUB7554642. Moreover, INSDC policy does

not require that data be removed in the case of erroneous submission,⁷ NIH refused to remove Submission ID SUB7554642 when the Wuhan University initially claimed that it had been submitted in error.

On June 23, 2021, in statements to reporters, the NIH's Renate Myles wrote that researchers who submit data to the SRA hold rights to such data, implied that the researchers' rights include having the data removed from the SRA. Myles wrote, "The requestor indicated the sequence information had been updated, was being submitted to another database and wanted the data removed from the SRA to avoid version control issues."

By contrast, the INSDC's written statement on data sharing during COVID-19 actually encourages submissions to multiple databases. "In cases where scientists have already established submissions to other databases, these submissions *should continue in parallel* to the INSDC submission."⁸

4. **In off-the-record emails, an NIH official steered reporters toward *Washington Post* coverage of Bloom's paper, which was more favorable to the NIH, and away from a *New York Times* article due to its "tone."** NIH officials expressed concern about the "tone" of a *New York Times* article. For example, the NIH's Renate Myles wrote to a reporter at *The Hill*, "Off the record: we think this WaPo story does a good job characterizing the situation," and provided a link to *The Washington Post* article. Similarly, she advised a reporter for *ABC*, "Off the record: the WaPo story is much more even-keeled than the NYT piece" and forwarded a link to her favored article.
5. **NIH Director Francis Collins personally reviewed and cleared the response to a reporter's FOIA request related to the sequence deletions.** "The FOIA Office had no objections to sharing the unredacted version of this response with Dr. Brennan and Dr. Collins," wrote an NIH official while reviewing a FOIA response. "Also, they will both be involved in clearing the final response before it is sent to the requestor."
6. **Although NIH still has copies of all "withdrawn" sequences "for preservation purposes," it refused to examine them in a transparent process, as proposed by Professor Jesse Bloom.** Bloom proposed an open scientific collaboration to determine whether any of the preserved data might help explain how the pandemic began. In October 2021, Bloom contacted NIH to discuss cooperating to analyze the deleted sequences. However, the NIH's Steve Sherry dismissed the idea claiming, "As you know, when data sets are withdrawn from the database, that status does not permit use for further analyses."
7. **Bloom pressed the NIH about another, separate set of deletions being examined by "an investigative entity."** Bloom discovered a separate set of deleted sequences that had "reappeared" without explanation. A week after Sherry dismissed his proposed collaboration, Bloom wrote to Sherry again asking questions about what he called the "puzzling" reappearance of another previously unreported deletion of pangolin coronavirus sequences removed at the request of South China Agricultural University.

⁷ INSDC, "International Nucleotide Sequence Database Collaboration Policy," which in relevant part at ¶ 3 provides that "erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number" (emphasis added).

<https://www.insdc.org/policy.html>

⁸ INSDC, "Statement on SARS-CoV-2 sequence data sharing during COVID-19" (emphasis added).

[https://www.insdc.org/sites/insdc.org/files/documents/INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19.pdf](https://www.insdc.org/sites/insdc.org/files/documents/INSDC%20Statement%20on%20SARS-CoV-2%20sequence%20data%20sharing%20during%20COVID-19.pdf)

“To understand why they reappeared over a year after being deleted,” Bloom wrote, “an investigative entity sent a request to NLM/NIH for all correspondence related to these accessions[.]” Bloom questioned Sherry’s “previous explanation ... that once datasets are removed a submitter’s request, they are only restored if the submitter requests that.”

Bloom claimed that NIH had provided the “investigative entity” no evidence that the submitters in China had requested the data be restored. It is unclear whether Sherry answered Bloom’s questions about whether: (1) the submitters in China in fact asked to restore the sequences and NIH withheld that request from the “investigative entity,” or (2) the sequences were restored without such a request and if so, why.

The Documents

By November 17, 2021, NIH had failed to comply with Empower Oversight’s FOIA request from the previous July. Hence, Empower Oversight sued NIH in the United States District Court for the Eastern District of Virginia to compel its compliance with FOIA and to obtain the documents described in the July 14th FOIA request.

NIH’s FOIA staff appears to have made significant errors when searching for responsive records (missing documents that should have been found and produced) and when reviewing records for FOIA exemptions (thus, redacting content that should not have been redacted). However, the few documents that NIH has produced thus far contain significant new information that is outlined below.

The [entire 238-page cache of emails](#) is available for download. Below is a detailed description of what they show.

According to these emails, a researcher submitted genetic sequences to NIH for uploading to the SRA and then asked NIH to remove them. Specifically, the records show that the researcher tried unsuccessfully to get NIH to remove the sequences in early June 2020. Later that month, the researcher successfully persuaded NIH to remove the sequences, after he changed his rationale for the removal. Interestingly, the researcher’s first rationale for removal was compliant with NIH’s conditions for removal, but his latter rationale was not.

A year later, Professor Jesse Bloom discovered that public access to the sequences on the SRA had been removed and contacted NIH in June 2021 to discuss the matter. As Bloom explained in an email to the NIH, the gene sequences may help understand how the pandemic began.

NIH Director Francis Collins responded, “This is truly intriguing. I’ll be interested in [NIH official Steve Sherry’s] thoughts about the deleted SRA entries and whether there is any way to recover information about how that happened.”

Bloom later published a preprint on these removed virus sequences which generated several media stories, and an immediate reaction within NIH. Subsequently, Bloom tried to collaborate with the NIH on an analysis of the deleted sequences but was rebuffed by the NIH.

A Chinese researcher—whose identity was hidden by NIH in the documents produced through FOIA—submitted virus sequences to the NIH’s Sequence Read Archive (SRA) on March 17, 2020. According to a later story in *The New York Times*, the Chinese researcher’s name was Ben Hu, at Wuhan University.⁹ This submission was given the submission identification SUB7147304 and the reference PRJNA612766. The next day, the submitter of

⁹ “Those Virus Sequences That Were Suddenly Deleted? They’re Back,” *The New York Times* (Jul 30, 2021). <https://www.nytimes.com/2021/07/30/science/coronavirus-sequences-lab-leak.html>

SUB7147304 contacted the NIH to complain about an inability to download the data. The NIH responded that there was a delay in processing, but that the data was available.

On June 5, 2020, the Wuhan University researcher made an additional submission, which was give the submission identification SUB7554642 and the reference PRJNA637497.

```
bioprojecthelp at ncbi.nlm.nih.gov
Fri Jun 5 08:01:17 EDT 2020

Dear (b) (6)

This is an automatic acknowledgment that your submission:

SubmissionID: SUB7554642
BioProject ID: PRJNA637497
Title:

has been successfully registered with the BioProject database. After review
by the database staff, your project information will be accessible with the
following link, usually within a few days of the
release date that you set (or the release of linked data, whichever is
first):
```

```
http://www.ncbi.nlm.nih.gov/bioproject/637497

Please use the BioProject ID PRJNA637497 with your correspondence and your
data submissions.

Send questions to bioprojecthelp at ncbi.nlm.nih.gov, and include the
BioProject ID and organism name.

Regards,

NCBI BioProject Submissions Staff
Bethesda, Maryland USA
*****
(301) 496-2475
(301) 480-2918 (Fax)
bioprojecthelp at ncbi.nlm.nih.gov (for BioProject questions/replies)
info at ncbi.nlm.nih.gov (for general questions regarding NCBI)
*****
```

Later that day, the submitter asked to retract the submission, claiming unspecified error.

From: (b) (6)
Received: Fri Jun 05 2020 21:45:04 GMT-0400 (Eastern Daylight Time)
To: Bioproject Support <bioprojecthelp@ncbi.nlm.nih.gov>;
Subject: retract BioProject

Dear Mr/Ms,

I want to retract a submission, and the BioProject ID is PRJNA637497.
I'm sorry for my wrong submitting. Thank you for your help.

Regards

(b) (6)

The NIH replied that it prefers to edit or replace submissions, rather than delete or remove them.

From: NLM Support <nlm-support@nlm.nih.gov>;
Received: Mon Jun 08 2020 13:36:22 GMT-0400 (Eastern Daylight Time)
To: (b) (6)
Subject: Re: case #CAS-550133-G858X0: retract BioProject TRACKING:00030000004630

Dear (b) (6)

Thank you for your email. We prefer to edit an existing BioProject or change its status to "replaced by" a new BioProject, rather than delete. If you submitted another BioProject to replace this one, please provide the BioProject ID for that project and we will set the status of this project to "replaced by" the desired one.

We have implemented a new capability that allows submitters to view the current content of a BioProject and make minor edits, including updating the title and description, and changing the release date. Please go to the submission portal and click on "Manage Data" where you can access your BioProject. Click on the BioProject accession in the left ("Accession") column and you will have the opportunity to make the desired change. The updates will be processed automatically and the page should refresh with the edited information within a few minutes (typically seconds). You will then be able to make additional changes, if needed.

If you need to make changes in other fields, please email the desired changes and we will edit for you. If you do not plan to use this BioProject or submit a replacement, we can delete it.

If you have other comments or questions, please reply to bioprojecthelp@ncbi.nlm.nih.gov.

Best regards,

(b) (6)
BioProject Curation Staff

* PLEASE DO NOT MODIFY THE SUBJECT LINE OF THIS EMAIL WHEN RESPONDING TO ENSURE CORRECT TRACKING *

Case Information:
Case #: CAS-550133-G858X0
Customer Name: (b) (6)
Customer Email: (b) (6)
Case Created: 2020-06-06T01:45:32Z

Summary: retract BioProject

Details:

Dear Mr/Ms,

I want to retract a submission, and the BioProject ID is PRJNA637497. I'm sorry for my wrong submitting. Thank you for your help.

Regards

(b) (6)

On June 15, the Wuhan University requestor submitted a second request to withdraw a related genetic sequence, citing submission of the data to another database.

From: [REDACTED] (b) (6)
Received: Mon Jun 15 2020 23:10:41 GMT-0400 (Eastern Daylight Time)
To: NLM/NCBI List sra <sra@ncbi.nlm.nih.gov>; SRA Support <sra@ncbi.nlm.nih.gov>;
Subject: Re: SUB7554642/subs/sra/SUB7554642/overview

Dear Mr/Ms,

Recently, I found that it's hard to visit my submitted SRA data, and it would also be very difficult for me to update the data. I have submitted an updated version of this SRA data to another website, so I want to withdraw the old one at NCBI in order to avoid the data version issue. The Submission ID is SUB7147304. I would appreciate your help.

Best regard,

[REDACTED] (b) (6)

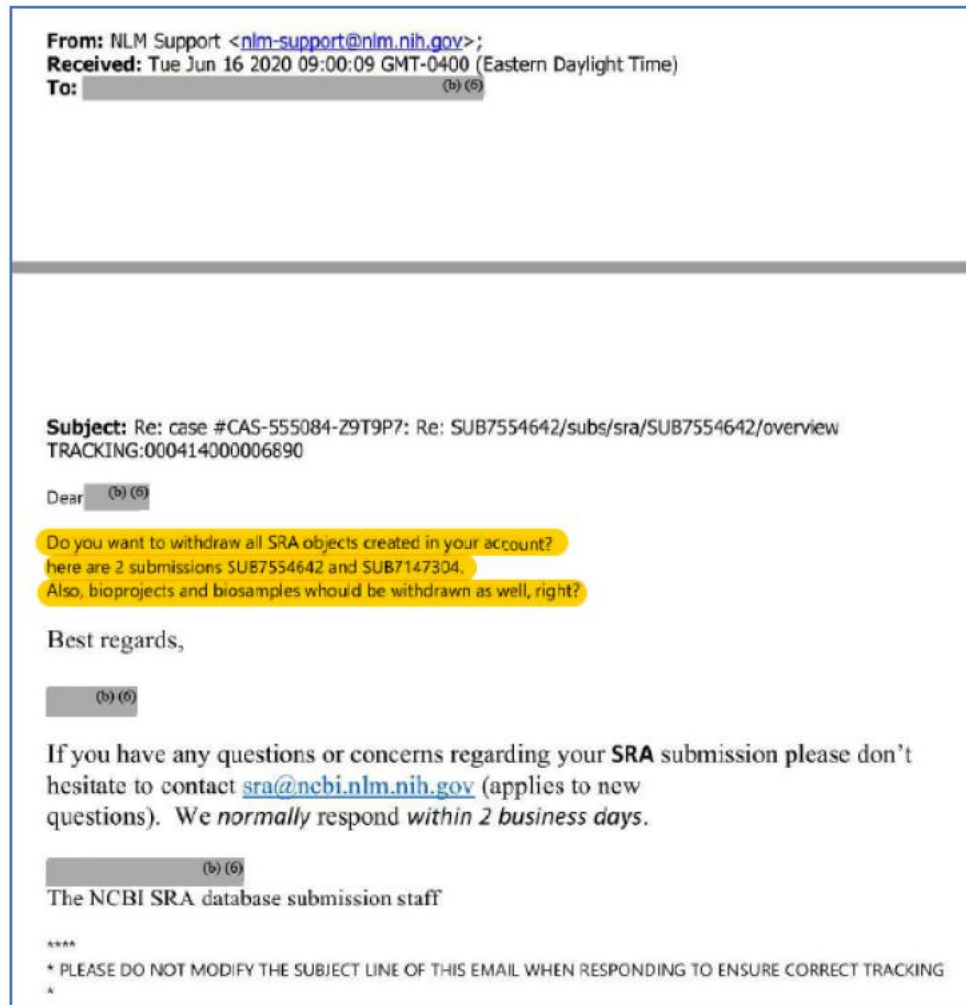
Summary: Re: SUB7554642/subs/sra/SUB7554642/overview

Details:
Dear Mr/Ms,

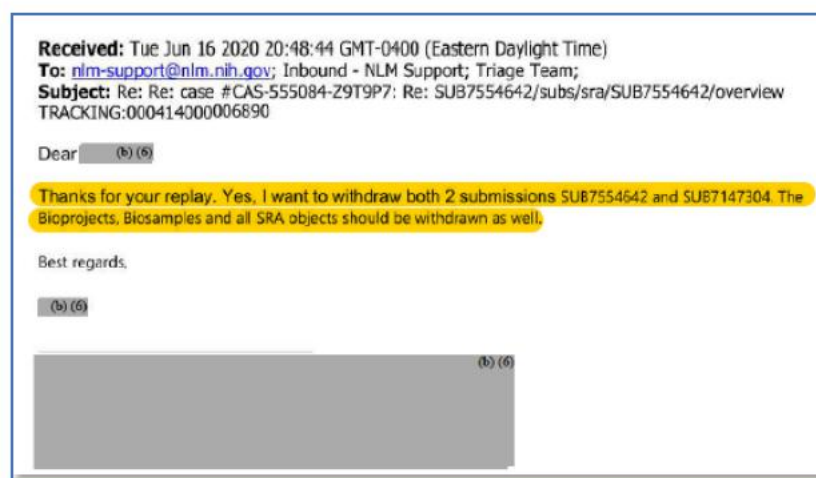
Recently, I found that it's hard to visit my submitted SRA data, and it would also be very difficult for me to update the data. I have submitted an updated version of this SRA data to another website, so I want to withdraw the old one at NCBI in order to avoid the data version issue. The Submission ID is XXXX. I would appreciate your help.

Best regard,
—
[Submitter]
Wuhan University

The next day, NIH agreed to the request, and sought clarification whether the Wuhan University researcher also wanted NIH to delete Submission ID is SUB7554642, which NIH had refused to remove a week prior.



The Wuhan University researcher responded that he/she wanted both submissions as well as all related bioprojects and biosamples removed.



Dear [SRA],

Thanks for your replay. Yes, I want to withdraw both 2 submissions XXXXX and YYYYY. The Bioprojects, Biosamples and all SRA objects should be withdrawn as well.

Best regards,

[Submitter]
Wuhan University

And, NIH replied that his/her request for removal had been accomplished.

From: NLM Support <nlm-support@nlm.nih.gov>;
Received: Wed Jun 17 2020 12:58:07 GMT-0400 (Eastern Daylight Time)
To: (b) (6)
Subject: Re: Re: case #CAS-555084-Z9T9P7: Re: SUB7554642/subs/sra/SUB7554642/overview
TRACKING:000414000006890

Hi (b) (6)

I had withdrawn everything.

Best regards,

(b) (6)

If you have any questions or concerns regarding your **SRA** submission please don't hesitate to contact sra@ncbi.nlm.nih.gov (applies to new questions). We *normally* respond *within 2 business days*.

(b) (6)
The NCBI SRA database submission staff

Around June 18, 2021, Professor Jesse Bloom of the Fred Hutchinson Cancer Research Center contacted officials at the NIH to inform them that he had identified a data set of early Wuhan virus sequences that been deleted from the SRA. Bloom wrote that it would be worthwhile to analyze this data and attached a preprint he had done looking into it. Bloom wrote to Francis Collins, Steve Sherry, and Anthony Fauci: "Anyway, I hope overall this can

be a good opportunity for the NIH to take the lead by using its remarkable data archives to make progress in resolving some of the important questions about the virus's origins,”

From: Bloom PhD, Jesse D <jbloom@fredhutch.org>
Sent: Friday, June 18, 2021 7:00 PM
To: Collins, Francis (NIH/OD) [E] (b) (6); Sherry, Steve (NIH/NLM/NCBI) [E] (b) (6); Fauci, Anthony (NIH/NIAID) [E] (b) (6)
Subject: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Hi Francis, Stephen, and Toni,

I'm just writing to give you a heads up that I identified a data set of early Wuhan SARS-CoV-2 sequences that has been deleted from the NIH's Sequence Read Archive (SRA). I was able to recover the deleted files from the Google Cloud and analyze the sequences, and have attached a pre-print on the analysis that I just submitted for posting by *bioRxiv*.

Since SARS-CoV-2 origins and early spread has become a hot-button topic, I wanted to give you a heads up. I made sure to emphasize in the discussion that the SRA has many sequences and so isn't in a position to vet all deletions. Nonetheless, I think it would be highly worthwhile to do a comprehensive analysis of SRA (and other NIH) data that might be relevant to this topic that could have been deleted or otherwise overlooked. If I can be of any assistance, let me know.

I have been running a pipeline to identify additional deleted SRA data using various heuristics including those in described in the attached pre-print, but have not yet completed the analysis enough to know the extent that the data I have recovered is relevant to SARS-CoV-2's origins or early spread. But as I mention in the pre-print, there are two known SRR deletions that are worth looking at. I definitely think it would be good to perform a SRA side search as well, since that will obviously be easier and more efficient, and could identify deleted data not on the cloud.

Anyway, I hope overall this can be a good opportunity for the NIH to take the lead by using its remarkable data archives to make progress in resolving some of the important questions about the virus's origins.

Thanks,
Jesse

Jesse Bloom
Associate Professor, Fred Hutch Cancer Research Center
Affiliate Associate Professor, Genome Sciences & Microbiology, University of Washington
Investigator, Howard Hughes Medical Institute

Francis Collins responded, “This is truly intriguing. I’ll be interested in Steve’s thoughts about the deleted SRA entries and whether there is any way to recover information about how that happened.”

From: Brennan, Patti (NIH/NLM) [E] (b) (6)
Date: Saturday, June 19, 2021 at 5:05 AM
To: Collins, Francis (NIH/OD) [E] (b) (6), Pruitt, Kim (NIH/NLM/NCBI) [E] (b) (6)
Cc: Fauci, Anthony (NIH/NIAID) [E] (b) (6), Bloom PhD, Jesse D. <jbloom@fredhutch.org>
Subject: Re: URGENT: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Good morning colleagues

I just spoke to Francis and Steve Sherry. Steve is investigating the situation and will brief Francis later this morning

Patti

Patricia Flatley Brennan, RN, PhD
Director, National Library of Medicine
National Institutes of Health
US Department of Health and Human Services
Telework Hours 830-5 and by appt

From: Collins, Francis (NIH/OD) [E] (b) (6)
Sent: Saturday, June 19, 2021 7:37:59 AM
To: Pruitt, Kim (NIH/NLM/NCB) [E] (b) (6)
Cc: Brennan, Patti (NIH/NLM) [E] (b) (6); Fauci, Anthony (NIH/NIAID) [E] (b) (6); Bloom PhD, Jesse D <jbloom@fredhutch.org>
Subject: URGENT: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Hi Kim,

See note below and the attached rather stunning preprint. (b) (5)
(b) (5). I got an “out of office” from Steve saying he was gone until June 28. (b) (5)
(b) (5)
(b) (5)

Please let me know right away what can be learned about this.

Francis

From: Collins, Francis (NIH/OD) [E]
Sent: Friday, June 18, 2021 10:01 PM
To: Bloom PhD, Jesse D <jbloom@fredhutch.org>; Sherry, Steve (NIH/NLM/NCBI) [E] (b) (6); Fauci, Anthony (NIH/NIAID) [E] (b) (6)
Subject: RE: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Dear Jesse,

This is truly intriguing. I’ll be interested in Steve’s thoughts about the deleted SRA entries and whether there is any way to recover information about how that happened.

Francis

The following day, on June 19, Kim Pruitt, Senior Scientist and Chief of the NIH's Information and Engineering Branch, emailed her colleagues Steve Sherry and Patti Brennan, "The only way data is removed from SRA (per SOP) is if a submitter notifies us that the submission was in error. We would not delete data ourselves. Only submitters have that authority over their data."

From: Brennan, Patti (NIH/NLM) [E]
Sent: Sat, 19 Jun 2021 09:34:52 -0400
To: Pruitt, Kim (NIH/NLM/NCBI) [E]; Sherry, Steve (NIH/NLM/NCBI) [E]
Subject: Re: URGENT: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Thanks that is a helpful update- (b) (5)

Patti
Patricia Flatley Brennan, RN, PhD
Director, National Library of Medicine
National Institutes of Health
US Department of Health and Human Services
Telework Hours 830-5 and by appt

From: Pruitt, Kim (NIH/NLM/NCBI) [E] (b) (6)
Sent: Saturday, June 19, 2021 9:13:21 AM
To: Brennan, Patti (NIH/NLM) [E] (b) (6); Sherry, Steve (NIH/NLM/NCBI) [E] (b) (6)
Subject: RE: URGENT: SARS-CoV-2 data deleted from the NIH/NCBI SRA

Patti,
We are trying to find a way to contact people who work on SRA and trying to search Dynamics for prior communications. We don't have home phone numbers, trying to contact others who might possibly have that info.

The only way data is removed from SRA (per SOP) is if a submitter notifies us that the submission was in error. We would not delete data ourselves. Only submitters have that authority over their data.

Kim

*Kim D. Pruitt, Ph.D
Senior Scientist
Chief, Information Engineering Branch, NCBI/NLM/NIH*

*Telework hours: 8:30 – 5:00
Phone: (b) (6)*

*45 Center Drive
Building 45 Room 5AN44A
Bethesda, MD 20892-6511*

From: Brennan, Patti (NIH/NLM) [E] (b) (6)

On June 20, 2021, NIH Director Francis Collins and Anthony Fauci, the Director of the National Institute of Allergy and Infectious Diseases, hosted a Zoom call with Jesse Bloom and other research academics, including Kristian Andersen, Robert F. Garry, Trevor Bedford, Sergei Pond, and Rasmus Nielsen.

Prior to the call, Director Collins shared a suggested agenda with the invitees.

From: Francis Collins (b) (6)
Date: Saturday, June 19, 2021 at 11:07 AM
To: (b) (6), "Garry, Robert F" <rfgarry@TULANE.EDU>, "spond@temple.edu" <spond@temple.edu>, "rasmus_nielsen@berkeley.edu" <rasmus_nielsen@berkeley.edu>, "tbedford@fredhutch.org" <tbedford@fredhutch.org>
Cc: Anthony Fauci (b) (6), "Bloom PhD, Jesse D" <jbloom@fredhutch.org>, "Embry, Alan (NIH/NIAID) [E]" (b) (6), "Tabak, Lawrence (NIH/OD) [E]" (b) (6)
Subject: URGENT: Seeking your expert advice

Hi Kristian, Bob, Sergei, Rasmus, and Trevor,

Tony Fauci and I would like to get your advice on the interpretation and significance of a preprint that Jesse Bloom has just submitted to BioRxiv (attached). As you will see, through some clever sleuthing, Jesse has been able to discover 13 sequences of SARS-CoV-2 spike protein that were deposited (and then deleted) from the SRA by a Chinese investigator at Wuhan University. The sequences are incomplete but interesting, in that they appear to represent a slightly closer relationship to RaTG13 than the prior root of the phylogenetic tree.

Would you be willing to have a close read of the paper and then join a conference call with Jesse, Tony, and me tomorrow (Sunday 6/20) at 3 PM EDT? Steve Sherry of NCBI will also join – he has been digging out information about how these reads were removed from SRA by a request from the submitter, and assessing whether there might have been any other similar requests in early 2020.

Let me know if you can be available.

Thanks, Francis

The following day, on June 21, Trevor Bedford sent an email to the group stating that the newly recovered data seem to support that the idea that the pandemic began outside the Huanan market in Wuhan and that the matter must be analyzed properly.¹⁰

From: Trevor Bedford
Sent: Mon, 21 Jun 2021 13:19:56 -0400
To: Collins, Francis (NIH/OD) [E]
Cc: Tabak, Lawrence (NIH/OD) [E]; Fauci, Anthony (NIH/NIAID) [E]; Embry, Alan (NIH/NIAID) [E]; Sherry, Steve (NIH/NLM/NCBI) [E]; Kristian G. Andersen; Bloom PhD, Jesse D; Garry, Robert F; rasmus_nielsen@berkeley.edu; spond
Subject: Re: URGENT: Seeking your expert advice

Hi all,

My apologies for missing the meeting yesterday. I don't generally check my @fredhutch.org email address on weekends and I missed this entirely.

I'm not sure what the consensus was on the call, but my general take is as follows: Although there were clearly confirmed cases that were non-market associated in early December, the large market outbreak had remained a major datapoint for me in a zoonotic scenario, as emergence outside the market would require a very early transmission chain make it to the market and be amplified (not impossible, but less parsimonious).

However, these new sequences add to phylogenetic evidence that the root of the SARS-CoV-2 phylogeny may well lie in lineage A rather than lineage B and support a root that's the outside the market.

Rasmus, Sergei and Jesse have all worked on this rooting issue. If we could have confidence that the root of the phylogeny does not match with market-associated genomes this would be strong evidence for me that the market is a secondary foci and not the site of emergence. I view this rooting issue as highly important to analyze properly and to determine uncertainty between different root locations.

Best,
- Trevor

¹⁰ On social media, Bedford tweeted days later, "As I've said before, I believe both zoonosis and lab leak to be plausible hypotheses for COVID origins. I'm not pushing any narrative, just trying to figure out what's going on with this particular datapoint."
<https://twitter.com/trvr/status/1408080730064703493?s=20&t=5AgngTRBsUWxUG-tYMSNsw>

On June 23, NIH began to receive requests for comment about the Jesse Bloom preprint. When a reporter from *The Hill* contacted NIH, Renate Myles emailed him a prepared NIH statement that began, “Thanks for checking with us. The below statement is attributable to NIH generally.” Myles added, “The requestor indicated the sequence information had been updated, was being submitted to another database and wanted the data removed from the SRA to avoid version control issues.”

This statement from Myles seems at odds with a document on data policy that states that scientists should submit to multiple databases in parallel: “In cases where scientists have already established submissions to other databases, these submissions should continue in parallel to the INSDC submission.”¹¹ However, instead of providing this policy, Myles provided the journalist a link to a document titled, “INSDC Status Document.”¹² This document describes five categories status that data may have (public, confidential, suppressed, replaced, or killed) as well as the “causes” and “implications” of each status.

Myles also wrote, “Off the record: we think this WaPo story does a good job characterizing the situation,” and provided a link to the story.

From: [Myles, Renate \(NIH/OD\) \[E\]](#)
To: [Nathaniel Weiskol](#)
Cc: [Fine, Amanda \(NIH/OD\) \[E\]](#); [Emma Wojtowicz](#)
Subject: RE: Statement on database deletion?
Date: Wednesday, June 23, 2021 5:14:00 PM

Hi Nathaniel:


Thanks for checking with us. The below statement is attributable to NIH generally. Off the record: we think this WaPo story does a good job characterizing the situation:
https://www.washingtonpost.com/health/coronavirus-origin-nih-gene-sequence-deletion/2021/06/23/186e87d0-d437-11eb-a53a-3b5450fdca7a_story.html

NIH is aware of Dr. Bloom’s preprint submission. Staff at the National Library of Medicine (NLM), which hosts the Sequence Read Archive (SRA), have reviewed the submitting investigator’s request to withdraw the data. These SARS-CoV-2 sequences were submitted for posting in SRA in March 2020 and subsequently requested to be withdrawn by the submitting investigator in June 2020. The requestor indicated the sequence information had been updated, was being submitted to another database, and wanted the data removed from SRA to avoid version control issues. The submitting investigator published relevant information about these sequences [by preprint in March, 2020](#) and in a [journal in June, 2020](#). Submitting investigators hold the rights to their data and can request withdrawal of the data.

Currently, NLM has no plans to change the policy that recognizes submitters rights to their own data and the right to petition that their data be withdrawn from the SRA. The National Center for Biotechnology Information (NCBI), part of the NLM that manages the database, is the U.S. participating member of the International Nucleotide Sequence Database Collaboration (INSDC), which provides guidelines for withdrawing data: <http://www.insdc.org/documents/insdc-status-document>. NLM/NCBI can’t speculate on motive beyond a submitter’s stated intentions.

Thanks,
Renate

Renate Myles, MBA
Acting Associate Director for Communications and Public Liaison
Acting Director, Office of Communications and Public Liaison
National Institutes of Health
Tel: (b) (6)



¹¹ INSDC, “Statement on SARS-CoV-2 sequence data sharing during COVID-19.”
[https://www.insdc.org/sites/insdc.org/files/documents/INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19.pdf](https://www.insdc.org/sites/insdc.org/files/documents/INSDC%20Statement%20on%20SARS-CoV-2%20sequence%20data%20sharing%20during%20COVID-19.pdf)

¹² INSDC Status Document
<https://www.insdc.org/documents/insdc-status-document>

Later in the day, an NIH official emails about an article in *The Washington Post*: “NIH doesn’t come out badly, unless you read this paragraph as referring to NIH as well, or instead of, the scientist who withdrew the data.”¹³

“A New York Times article is also out, I don’t like its tone,” another official responds.

From: Pruitt, Kim (NIH/NLM/NCBI) [E]
Sent: Wed, 23 Jun 2021 17:40:33 -0400
To: Coleman, Janet (NIH/NLM/NCBI) [C]; Mizrachi, Ilene (NIH/NLM/NCBI) [E]; Skripchenko, Yuriy (NIH/NLM/NCBI) [C]; Brister, James (NIH/NLM/NCBI) [E]
Cc: Hicks, Denise (NIH/NLM/NCBI) [C]; Fleischmann, Lydia (NIH/NLM/NCBI) [C]; Trawick, Bart (NIH/NLM/NCBI) [E]; Sherry, Steve (NIH/NLM/NCBI) [E]
Subject: RE: SRA was contacted by NY times reporter

A New York Times article is also out. I don't like its tone.

<https://www.nytimes.com/2021/06/23/science/coronavirus-sequences.html>

Kim D. Pruitt, Ph.D
Senior Scientist
Chief, Information Engineering Branch, NCBI/NLM/NIH

Telework hours: 8:30 – 5:00
Phone: (b) (6)

45 Center Drive
Building 45 Room 5AN44A
Bethesda, MD 20892-6511

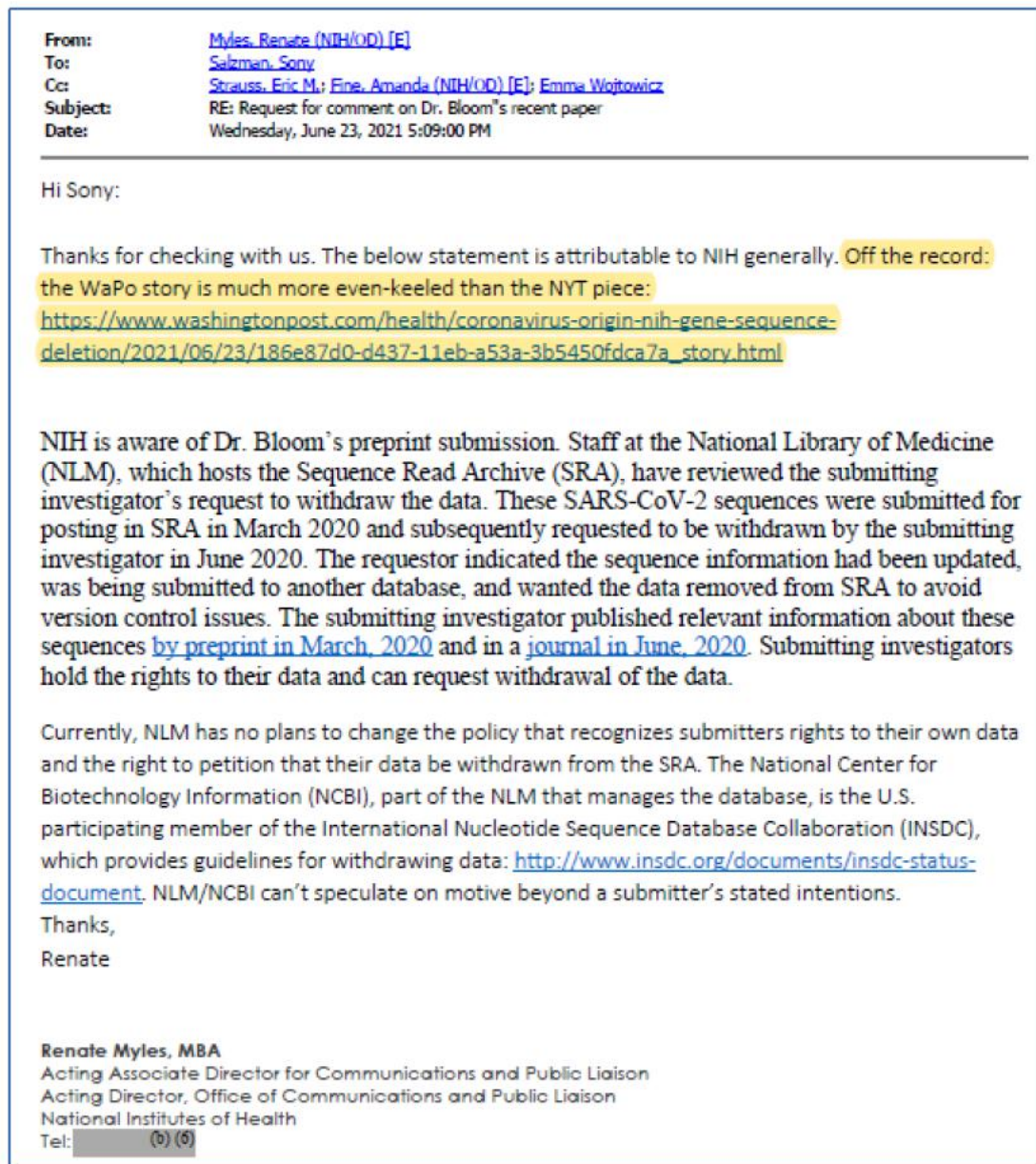
From: Coleman, Janet (NIH/NLM/NCBI) [C] (b) (6)
Sent: Wednesday, June 23, 2021 5:36 PM
To: Mizrachi, Ilene (NIH/NLM/NCBI) [E] (b) (6); Skripchenko, Yuriy (NIH/NLM/NCBI) [C] (b) (6); Pruitt, Kim (NIH/NLM/NCBI) [E] (b) (6); Brister, James (NIH/NLM/NCBI) [E] (b) (6)
Cc: Hicks, Denise (NIH/NLM/NCBI) [C] (b) (6); Fleischmann, Lydia (NIH/NLM/NCBI) [C] (b) (6); Trawick, Bart (NIH/NLM/NCBI) [E] (b) (6); Sherry, Steve (NIH/NLM/NCBI) [E] (b) (6)
Subject: Re: SRA was contacted by NY times reporter

For anyone who didn't yet see it, The Bloom study is on the top section of the online Washington Post :
https://www.washingtonpost.com/health/coronavirus-origin-nih-gene-sequence-deletion/2021/06/23/186e87d0-d437-11eb-a53a-3b5450fdca7a_story.html

NIH doesn't come out badly, unless you read this paragraph as referring to NIH as well as, or instead of, the scientist who withdrew the data: “Bloom said in an email to The Washington Post that he was not accusing the NIH of wrongdoing. But Bloom’s online paper suggests the deletion of data violates scientific norms and the code of trust essential to science”

¹³ “Seattle scientist digs up deleted coronavirus genetic data, adding fuel to the covid origin debate,” *The Washington Post* (Jun 23, 2021). https://www.washingtonpost.com/health/coronavirus-origin-nih-gene-sequence-deletion/2021/06/23/186e87d0-d437-11eb-a53a-3b5450fdca7a_story.html

And, to an ABC reporter, Myles characterizes *The New York Times* article as less “even-keeled” than *The Washington Post* article, and forwards a link to the article she favored.



On July 8, an NIH official began discussions about several FOIA requests that had been sent about the sequence deletions. The official's email states that the NIH FOIA office will determine redactions but that Francis Collins and Patricia Brennan will be involved.

According to the official's email, "The FOIA Office had no objections to sharing the unredacted version of this response with Dr. Brennan and Dr. Collins. Also, they will *both be involved in clearing the final response* before it is sent to the requestor."

From: Benson, Dennis (NIH/NLM/NCBI) [E]
Sent: Thu, 8 Jul 2021 16:25:12 -0400
To: Sherry, Steve (NIH/NLM/NCBI) [E]
Cc: Coleman, Janet (NIH/NLM/NCBI) [C]
Subject: Proposed response to FOIA request 56587 (Jon Cohen, Science)
Attachments: 56587 FOIA request 07-08-21.docx

Hi Steve – Janet compiled all the e-mail correspondence related to BioProject PRJNA612766 which was the project Jesse Bloom referred to in his paper. We reviewed the sequence and the contents of all the correspondence which was provided by Eric and Ilene.

BioProject PRJNA637497 was also mentioned in the correspondence associated with PRJNA612766 and therefore falls within the scope of Request 56597 from Jon Cohen, Science Magazine: "email correspondence with researche(r)s who requested the following data removed from the NCBI SRA Database".

Please supply me with the e-mail correspondence with researches who requested the following data removed from the NCBI SRA Database, which Jesse Bloom has described in this preprint: <https://www.biorxiv.org/content/10.1101/2021.06.18.449051v1>

The FOIA Office is in negotiations with the other six requesters to determine if they will be satisfied in limiting their requests to the same scope as the Cohen request.

The attachment contains no redactions. Redactions will be decided by the NHLBI FOIA Office in consultation with the NIH FOIA Office.

The FOIA Office had no objections to sharing the unredacted version of this response with Dr. Brennan and Dr. Collins. Also, they will be both be involved in clearing the final response before it is sent to the requestor.

Let me know if you have any questions.


Per Patti's request we can share this preliminary response with Diane Tuncer and Tara Mowery.


Dennis

On July 23, NIH officials shared a news article from China's *Xinhua* news agency in which Chinese officials attempted to explain why the data had been removed. However, Chinese officials claimed, "They found that the uploading address where the sequencing data

can be found *was deleted during the review* of the paper. Therefore, it was deemed unnecessary to keep their data in an NIH database.”

Chinese officials also accused Jesse Bloom of having “concocted the conspiracy theory that it was a coverup” and committing “a violation of scientific ethics.”

| | |
|--|--|
| <p>From: Tuncer, Diane (NIH/NLM) [E] Sent: Mon, 26 Jul 2021 11:54:02 -0400 To: Sherry, Steve (NIH/NLM/NCBI) [E]; Pruitt, Kim (NIH/NLM/NCBI) [E] Cc: Crutchman, Alise (NIH/NLM) [E] Subject: FW: FYI - News report from China's Xinhua News Agency Attachments: NLM SRA Deletion Media Responses 7.1.2021 New Responses.docx</p> <p>Hi Steve and Kim,</p> <p>We shared this news article with NIH OCPL too, and they [NIH OCPL] responded with the following information (see below). I'm also attaching the last set of QA (which we already sent to you last week).</p> <p>***</p>  <p>From: Tuncer, Diane (NIH/NLM) [E] Sent: Friday, July 23, 2021 8:46 AM To: Brennan, Patti (NIH/NLM) [E]; Sherry, Steve (NIH/NLM/NCBI) [E] Cc: Nurik, Jody (NIH/NLM) [E]; Pruitt, Kim (NIH/NLM/NCBI) [E]; Crutchman, Alise (NIH/NLM) [E] Subject: FYI - News report from China's Xinhua News Agency</p> <p>Thought you would be interested in the following news report from China's major news agency, Xinhua, in response to Dr. Bloom's paper. In the article, China's deputy head of the National Health Commission provides an explanation for the events that led to the request. See highlighted text below.</p> | <p style="text-align: center;">Claim that Chinese researchers hid coronavirus data defies scientific ethics</p> <p><i>Xinhua, July 23, 2021</i> Adjust font size: [icon]</p> <p>A staff member carries out testing at the inactivated COVID-19 vaccine quality inspection lab of Sinovac Life Sciences Co., Ltd. in Beijing, capital of China, Dec. 23, 2020. [Photo/Xinhua]</p> <p>A U.S. researcher claiming that China hid coronavirus sequences to thwart the tracing of virus origin is against scientific ethics, a Chinese official said Thursday.</p> <p>Last year, Chinese researchers published a research paper titled "Nanopore targeted sequencing for the accurate and comprehensive detection of coronavirus and other respiratory viruses" on the journal <i>Small</i>.</p> <p>Jesse Bloom is a computational biologist and specialist in viral evolution at the Fred Hutchinson Cancer Research Center in Seattle. Last month, he said that the coronavirus sequences in the study had been removed from the Sequence Read Archive, an online database run by the U.S. National Institutes of Health (NIH), at the request of Chinese researchers.</p> <p>Bloom said he was able to recover copies of the data stored on Google Cloud. "It therefore seems likely the sequences were deleted to obscure their existence" and "suggests a less than wholehearted effort to trace early spread of the epidemic," Bloom wrote in a preprint paper, not yet peer-reviewed by other scientists.</p> <p>Speaking at a press conference on the novel coronavirus origin-tracing, Zeng Yixin, deputy head of the National Health Commission, said that China investigated the claim after it was reported.</p> <p>The research paper is about a sequencing approach to help detect the coronavirus. According to Zeng, when the researchers submitted the paper last March, they needed to upload the sequencing results to prove their method.</p> <p>On June 9, 2020, the journal sent the sample paper ready to be published to the researchers. They found that the uploading address where the sequencing data can be found was deleted during the review of the paper. Therefore, it was deemed unnecessary to keep their data in an NIH database. On June 16, 2020, the Chinese team emailed NIH to remove the data, and NIH removed the data at the request.</p> <p>"The researcher has no need to hide or cover up and has no such subjective intention," Zeng said. Meanwhile, the researchers have uploaded the sequencing data, including 244 pieces of data from</p> |
|--|--|

| | |
|---|--|
| <p>61 samples, to the GSA database under China's National Genomics Data Center. The database is open to global users and anyone can make an inquiry.</p> <p>Zeng added that the earliest sampling time of the virus samples is on Jan. 30, 2020, which has been some time since the beginning of the epidemic. The information and research value that these sample sequencing can provide is very limited in the coronavirus origin tracing.</p> <p>Jesse Bloom did not get the confirmation from the Chinese researchers, did not understand the background of the data removal, and concocted the conspiracy theory claiming that it was a cover-up, Zeng said.</p> <p>He noted that Bloom's conspiracy theory has a bad influence on international public opinion, slandered Chinese researchers and hurt them. "It is not only a departure from science but also a violation of scientific ethics."</p> <p>During epidemics such as the COVID-19, the public pays attention to every word and action of scientists. Therefore, scientists should know their social responsibilities and not make arbitrary speculations, said Zeng, pointing out that Bloom's paper has been criticized by many scientists.</p> <p>Diane Tuncer, MPH Supervisory Writer/Editor Office of Communications and Public Liaison National Library of Medicine National Institutes of Health Mobile: [icon] [icon] [icon] [icon]</p>  | <p>General Response That Had Been Provided to the Media:</p> <p>Early in the pandemic, NIH and other federal agencies moved quickly to make COVID-19 open-access data and computational resources freely available to researchers. NIH's National Library of Medicine has a broad portfolio of open-access databases, including the Sequence Read Archive (SRA), the world's largest publicly available repository of high-throughput sequencing data. In the past year, SRA received approximately 2.4 million submissions of sequence data.</p> <p>SRA is managed by NLM's National Center for Biotechnology (NCBI), which is the U.S. participating member of the International Nucleotide Sequence Database Collaboration (INSDC) since 1987. NCBI follows the INSDC policies and guidelines for data submission and change requests, and collaborates with participating organizations on updating policies and guidelines as described in this 2018 article. The guidelines describe the criteria for which submitting researchers can request a change in data status (for example, if the data have been corrupted) and actions taken if the criteria are met.</p> <p>In March 2020, the SARS-CoV-2 sequences in question were submitted by a researcher at a China-based institution for posting in SRA. In June 2020, in response to a request by the same researcher, NCBI withdrew the sequences.</p> <p>NCBI has initiated an independent review of SRA processes and standard operating procedures to determine whether the appropriate steps were taken to assess this withdrawal request. Withdrawal makes the data undiscoverable but does not erase it. Per the INSDC guidelines, NCBI retains withdrawn data for the scientific record and for disaster recovery. Pending outcome of the review, NCBI will work with INSDC to assign the data to the appropriate status.</p> <p>The researcher from the China-based institution published relevant information about these sequences by preprint in March, 2020 and in a journal in June, 2020.</p> <p>Will NLM/NCBI change its policy about data removal?</p> <p>NLM/NCBI considers the policies and guidelines of the INSDC sound. NCBI has initiated an independent review of SRA processes and standard operating procedures to determine whether the appropriate steps were taken to assess this withdrawal request. Withdrawal makes the data undiscoverable but does not erase it. Per the INSDC guidelines, NCBI retains the data for the scientific record and for disaster recovery. Pending outcome of the review, NCBI will work with INSDC to assign the data to the appropriate status.</p> <p>Can you say anything about whether NIH is doing any analysis or examination to look for any other SARS-CoV-2 sequence data that has been deleted from that database?</p> <p>NLM/NCBI's analysis found that from January 2020 through June 2021 six institutions requested withdrawal of SARS-CoV-2 submission packages through NLM/NCBI services. This included one requested by a researcher at a China-based institution and the rest from researchers at institutions from other countries, predominantly the U.S. In addition, five institutions requested withdrawal of sequence data through INSDC partners which were replicated within the SRA. NCBI has initiated an independent review of SRA processes and standard operating procedures to determine whether the appropriate steps were taken to assess this withdrawal request.</p> |
|---|--|

On September 27, Bloom emailed Steve Sherry to ask if the NIH planned to do a detailed report of the sequence deletions and if he can help. “I wanted to reach out to you with a proposal to search all deleted deep sequencing datasets on the SRA for sequences that might be relevant to SARS-Cov-2,” Bloom wrote.

Bloom added that he had read a *Wall Street Journal* article¹⁴ that reported the information was still accessible. He also noted that the NIH had inaccurately told *The Washington Post* there were only eight deletions.

From: Bloom PhD, Jesse D <jbloom@fredhutch.org>
Sent: Monday, September 27, 2021 12:25 AM
To: Sherry, Steve (NIH/NLM/NCBI) [E] [REDACTED] (b) (6)
Subject: Proposal for searching all deleted/suppressed SRA datasets

Hi Steve,

Hope all is well.

I wanted to reach out to you with a proposal to search all deleted deep sequencing datasets on the SRA for sequences that might be relevant to SARS-CoV-2. Apologies if you also hear about this idea from others as I have been running it by various others for feedback too, but I figured maybe I should directly get in touch with you as well.

As you probably know, the question is whether any datasets might have been deleted or suppressed

that contained sequences relevant to SARS-CoV-2. These could either be viral sequences or sequences with just contamination from viral reads.

I have been able to build a list of 122,904 accessions (SRR, ERR, and DRR) that became “suppressed” (which includes both suppressed and killed in the terminology of the INSDC status document) between 2018-12-02 and 2021-08-10. For most of them, I’ve also been able assemble relevant metadata such as dates of status changes, number of reads, md5 checksums, and in some cases other information. From this information, I’ve been able to partially prioritize them. I downloaded and analyzed as many as are still available through the SRA or Google / Amazon cloud, which is unfortunately only 1829 of the 122,904. Of the remaining, based on the metadata I rank 565 as being of the highest priority, 2822 of medium priority, 29160 of moderate priority, and 88528 of lower priority. I am trying to obtain more of these datasets from other sources (there are a few organizations that download and store large amounts of SRA data), but I’m sure I will not be able to get many of them.

¹⁴ Amy Dockser Marcus and Drew Hinshaw, “After Covid-19 Data Is Deleted, NIH Reviews How Its Gene Archive Is Handled,” *The Wall Street Journal* (Sep 13, 2021).
<https://www.wsj.com/articles/after-covid-19-data-is-deleted-nih-reviews-how-its-gene-archive-is-handled-11631545490>

I read in the [Wall Street Journal article a few weeks ago](#) how the SRA keeps copies of all accessions even if they have been removed from public access. So my proposal is that we come up with some strategy to analyze all of these deleted accessions. I have scalable Snakemake pipelines that can process this number of sequences, first to identify those with SARS-CoV-2 reads, and then place those reads in a phylogenetic context to identify any more “ancestral” looking sequences. Here on the Hutch cluster I could process ~100,000 accessions in somewhere between 2-6 weeks depending on how much time is needed to transfer the files, and the pipelines should be relatively portable to run on another cluster if that is preferable.

I think that doing this type of analysis could be consistent with INSDC policy. For instance, the [main INSDC policy page](#) actually says that data submitted to the INSD will always remain permanently accessible. Although [the INSDC status page](#) conflictingly says in rare cases data can be killed, it still says there is no prior restraint on its use. Furthermore, the analysis would naturally discard all non-coronavirus reads, which would be the entirety of most datasets.

This approach could also help resolve some of the confusion about sequence deletions. I am now getting inquiries from congressional staff who are asking if the deletion of PRJNA612766 by Wuhan University was “proper” or should be investigated more. I explain that this question sort of misses the point: under INSDC status document policy, it is allowed for submitters to remove data. The correct question is not if the SRA was wrong to remove that project, but rather we are now doing everything we can to see if there is anything else of relevance now that we know these deletions can occur. I think this is especially important given the [recent revelations about the DARPA DEFUSE proposal](#) that highlight the possibility that there could be information relevant to SARS-CoV-2 that has been overlooked in the public discussion.

Finally, this could all be set up in a totally transparent way. For instance, the pipelines could be made available ahead of time along with the lists of accessions, and summary statistics could be output publicly. Therefore, in contrast to the brewing battles and investigations related to COVID-19 origins, for this part everything could be done totally transparently in a scientific framework that isn’t susceptible to speculation and doubt.

Anyway, let me know if you have any interest in chatting more about the possibility of some approach along these lines.

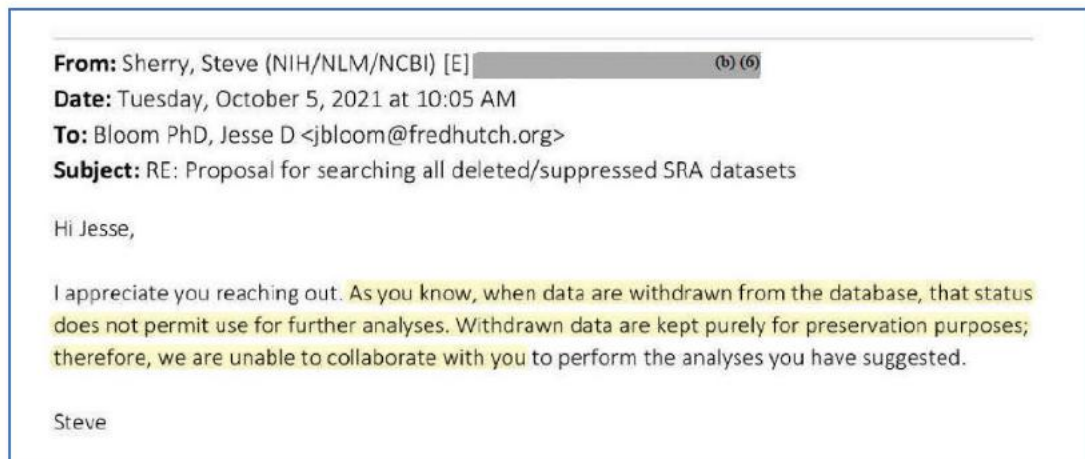
Also, I just wanted to mention that it turns out that I think there was an error in the statement the [NIH gave to the Washington Post](#) about the original Wuhan University deletions, where they said there were just 8 deletions and the rest were from submitters “predominantly in the US.” There were *at least* two other full-BioProject deletions that involves SARS-CoV-2-related reads: PRJNA637497 and PRJNA640246.

Thanks for considering all of this, and just let me know if there might be a chance to chat more.

--Jesse

On October 5, Sherry responded, refusing Bloom’s offer and claiming that that the agency was prohibited from analyzing the data even though the data is still accessible. Sherry wrote:

“As you know, when data sets are withdrawn from the database, that status does not permit use for further analyses. Withdrawn data are kept purely for preservation purposes.”



However, Sherry’s response conflicts with INSDC’s written policy. That policy states:

All database records submitted to the INSDC *will remain permanently accessible as part of the scientific record*. Corrections of errors and update of the records by authors are welcome and erroneous records *may be removed from the next database release*, but all will remain permanently accessible by accession number.¹⁵

A separate file called the “INSDC Status Document” also makes clear that records will always be available. Data may be “killed” if the INSDC violates a “confidential status” by a submitter, or if data were submitted by someone who is not the rightful owner of the data.¹⁶ If either of those indications have been met, the policy states:

Data are not directly available publicly from INSDC partners through any means. However, because the data will have been distributed previously as Public, the INSDC partners cannot exercise any control on the resultant use of the data by third parties.¹⁷

On October 12, Bloom contacted NIH again about more unexplained deletions by Chinese researchers, and their “puzzling” reappearance despite no evidence of further communication with the researcher. Bloom references “an investigative entity” without identifying it or explaining how he is familiar with what the NIH produced to that entity. It is unclear whether the entity could be the FBI and attempts to obtain further context from Professor Bloom by the publication date of this report were unsuccessful.

According to emails obtained through FOIA by *The Intercept*,¹⁸ the FBI contacted the NIH in May 2020 asking about grants made to EcoHealth Alliance, which is reported to have

¹⁵ International Nucleotide Sequence Database Collaboration Policy (emphasis added).
<https://www.insdc.org/policy.html>

¹⁶ INSDC Status Document
<https://www.insdc.org/documents/insdc-status-document>

¹⁷ *Id.*

¹⁸ “FBI Sought Document Related to U.S.-Funded Coronavirus Research in China,” *The Intercept* (Jan 20, 2022).
<https://theintercept.com/2022/01/20/coronavirus-research-china-ecohealth-fbi/>

collaborated with the Wuhan Institute of Virology.¹⁹ However, there have been no public reports confirming any FBI investigation into deleted gene sequences from the SRA. Without further context from Professor Bloom it is unclear what “investigative entity” he was referencing.

From: Bloom PhD, Jesse D
To: Sherry, Steve (NIH/NLM/NCBI) [E]
Subject: Question regarding two deleted and then restored deep sequencing runs
Date: Tuesday, October 12, 2021 1:20:00 AM

Hi Steve,

I'm writing to inquire about some more deleted deep sequencing runs from China on the SRA.

As you may know, two runs related to pangolin coronavirus sequences from China, SRR11119760 and SRR1119761 were deleted from the SRA on March-16-2020 by curator (b) (6) at the request of the submitter (b) (6) of South China Agricultural University under the stated rationale that they were accidental uploads unrelated to the project.

But a puzzling thing about these accessions is that they then re-appeared on the SRA over a year later, on or about June-16-2021.

To understand why they reappeared over a year after being deleted, an investigative entity sent a request to the NLM / NIH for all correspondence related to these accessions in the period spanning March of 2020 through June of 2021.

The documents that were provided in response to this request did not indicate any further correspondence between the submitters in China and the SRA after March of 2020 regarding these two samples.

We are therefore trying to understand the process and rationale by which the two deleted sequencing runs were again made available on the SRA. My understanding from your previous explanations is that once datasets are removed at a submitter's request, they are only restored if the submitter requests that. Yet the documents provided by NLM / NIH do not indicate that South China Agricultural University made any request to restore these accessions.

I am therefore wondering, which of the following is the case:

1. Was there in fact a request from South China Agricultural University to restore these sequences that was omitted from the documents provided by NLM / NIH?
2. Did the NCBI restore these sequences to public access without a request from South China Agricultural University? If so, what was the rationale and process for this restoration?

Thanks for your help in looking into this.

--Jesse

The documents provided thus far provide no information about how or whether Bloom's questions were answered.

Conclusion

These documents raise several questions that need further investigation to answer fully. Congress should press the NIH for answers on why it is stonewalling Senate inquiries and dragging its feet on basic transparency through FOIA. Most importantly, why has NIH refused to examine archival copies of deleted sequences in an open scientific process to determine whether any of that information might be able to shed light on the origins of the COVID-19 pandemic?

¹⁹ Paul Thacker, "Scientists Doing Dangerous Virus Research Cry Victim To Avoid Public Accountability," *The Disinformation Chronicle* (Jan 25, 2022).

<https://disinformationchronicle.substack.com/p/scientists-doing-dangerous-virus?s=r>