Notes and e-mails from Jesse Bloom related to my meeting with NIH on June-20-2021 after I sent them an advance copy of my pre-print about SARS-CoV-2 sequences that had been deleted from the NCBI Sequence Read Archive

About these notes: *I did not take notes at the time of the meeting. These notes were written almost 6 months later on January-14-2022 from memory and e-mail records. Immediately following the notes, I have included the limited e-mail records I have that discuss the content of the meeting with the NIH. I have not included other e-mails with the NIH from prior to the meeting, although I can provide those if requested.*

On June-18-2021, I completed a pre-print describing phylogenetic analyses of SARS-CoV-2 sequences that were deleted from the NIH's Sequence Read Archive, and submitted the pre-print to *bioRxiv*. The original copy of that pre-print is at https://www.biorxiv.org/content/10.1101/2021.06.18.449051v1.

Also on June-18-2021, I e-mailed the submitted pre-print to NIH Director Francis Collins, NIAID Director Anthony Fauci, and NCBI Director Steve Sherry. My goal in sending them the pre-print was to alert them about a high-profile and controversial topic, and perhaps initiate collaborative work with the NIH to determine if other SARS-CoV-2 related sequences had been deleted from NIH databases. That evening, Francis Collins replied that the findings were intriguing, and asked if the NCBI could determine more information. The next day, Steve Sherry located and sent a redacted version of the e-mail request by Wuhan University to delete the sequences (that e-mail is provided in a revised version of the pre-print I posted about a week later at https://www.biorxiv.org/content/10.1101/2021.06.18.449051v2). Steve Sherry also indicated he was asking the SRA to compile a list of all other withdrawn data for transparency.

On June-19-2021, Francis Collins sent an e-mail suggesting a Zoom meeting on Sunday (June-20-2021) to discuss the matter further. Francis Collins indicated he planned to invite two outside scientists, Kristian Andersen and Bob Garry. He also asked me if I would like to suggest any additional invitees, and I suggested Rasmus Nielsen, Sergei Pond, and Trevor Bedford—all of whom I consider experts on viral evolution. Trevor Bedford was unable to attend, but on Sunday (June-20-2021) we had a meeting that included myself, Kristian Andersen, Bob Garry, Rasmus Nielsen, and Sergei Pond, as well as the following NIH attendees: Francis Collins, Anthony Fauci, Steve Sherry, Lawrence Tabak, and Alan Embry.

The meeting began with Steve Sherry giving an overview of the INSDC policy that governs NCBI data, and a verbal explanation of how Wuhan University had requested these sequences be deleted which was allowable under NCBI policy. He also again indicated the NCBI was in the process of assembling a list of deleted datasets for transparency. I then started to give a summary of what I had found when I analyzed the deleted sequences.

At that point, the meeting became extremely contentious. Kristian Andersen strongly objected to my pre-print, and said he found it deeply troubling. I recall Kristian highlighting three major objections. First, Kristian contended that if the Chinese authors had decided to delete their data, it was unethical for me to analyze it further. Second, Kristian contended that the phylogenetic analyses in the pre-print were not interesting because there was nothing unusual about the phylogenetics of early SARS-CoV-2 sequences in Wuhan. This point was strongly disputed by Rasmus Nielsen, and my strongest memory of the meeting is Kristian and Rasmus

yelling at each other over Zoom: Rasmus stated that the inconsistency of date and outgroup rooting of SARS-CoV-2 in Wuhan (https://academic.oup.com/mbe/article/38/4/1537/6028993) is extremely puzzling and unusual, and Kristian stated that Rasmus did not understand viral phylogenetics and if he understood it the way Kristian did he would realize that there is nothing unusual about sequences from Wuhan. Finally, Kristian objected to my pre-print because he said that there was already intense criticism of scientists such as himself, he needed security outside his house, and my pre-print would fuel conspiratorial notions that China was hiding data and thereby lead to more criticism of scientists such as himself.

I do remember that Francis Collins, Anthony Fauci, Sergei Pond, and Bob Garry also contributed to the discussion, but I can no longer remember the precise details of what they said. The only specific point I remember is Anthony Fauci complaining about the wording in the Discussion of the pre-print where I said the Chinese researchers had "surreptitiously" deleted the sequences, pointing out that "surreptitious" was a loaded word and we couldn't know exactly why they had made the deletion.

Kristian Andersen then said that he was a screener at *bioRxiv*, and so he could delete the pre-print or revise it in a way that would leave no record that this had been done. I replied that although I appreciated the critiques, I was not going to delete the pre-print, and instead would go ahead with posting it. I also said that although I could imagine revising the pre-print, given the contentious nature of the meeting, I doubted it was appropriate to revise the pre-print in this context. At that point, both Anthony Fauci and Francis Collins clarified their views by each saying something to the effect: "Just for the record, I want to be clear that I never suggested you delete or revise the pre-print." Kristian continued to press the point that he could use his capacity as a screener at *bioRxiv* to upload a revised version of the pre-print. At that point, one of the NIH attendees (I think it was Francis Collins but I am not certain) said something to the effect: "Kristian, if he's already submitted the pre-print, it's better if you don't pressure him to revise it."

At that point, the meeting was basically over. I recall at the very end, Sergei Pond telling the NIH attendees that his strong advice was that they continue to keep an open mind about the origins and evolution of SARS-CoV-2, because there were things we still might not know.

Shortly after the meeting, I had brief e-mail exchanges with Sergei Pond and Rasmus Nielsen, which I have included as **Email 1** and **Email 2** at the end of this document.

The pre-print posted on June-22-2021. The next week, I did revise the pre-print (the revised version is at: https://www.biorxiv.org/content/10.1101/2021.06.18.449051v2). I made these revisions after soliciting feedback from another scientist (Stephen Goldstein) who posted his comments publicly on *bioRxiv*. You can see Stephen Goldstein's comments and a detailed item-by-item list of the revisions I made in the *bioRxiv* comments section at https://www.biorxiv.org/content/10.1101/2021.06.18.449051v1#comment-5435736918; I also summarized them in this Tweet thread https://twitter.com/jbloom_lab/status/1409945528612184065. I chose to revise the pre-print in this way as I think it is important to incorporate critiques, but also think that this should be done in an open fashion that ensures the original version remains available and the process that led to the revisions is transparent.

About a week after the pre-print posted, I received an e-mail from Jon Cohen, a reporter at *Science* magazine. In the e-mail, he said he had been told by an anonymous source that the NIH held a meeting where Francis Collins and Anthony Fauci asked me to not publish the pre-print. I clarified by providing a condensed version of the above description: there had been such a meeting, but that the only attendee who had explicitly raised the idea of deleting the pre-print was Kristian Andersen, and that Francis Collins and Anthony Fauci never made such a request. Jon Cohen's question and my reply are included below as **Email 3**.

I continue to believe it is important to investigate if additional data have been deleted from NIH databases. The NIH has not publicly released any detailed reports on this topic, but I continue to try to compile data on deletions and advocate with the NIH that it should be analyzed. See my proposal to Steve Sherry included below as **Email 4**.

Finally, I would like to point out that the INSDC status document (https://www.insdc.org/documents/insdc-status-document) specifically says that even if sequence data are deleted ("killed") after they have been publicly released, the INSDC does not exercise control on the use of that data by third parties that are able to obtain it. For this reason, I am confident that I was within my scientific rights to analyze the deleted data despite the objection mentioned above by Kristian Andersen that it was inappropriate for me to analyze data if the Chinese authors had tried to delete.

I am happy to provide any other e-mails related to the pre-print or meeting upon request.

Subject: Re: Moral support
Date: Sunday, June 20, 2021 at 3:52:39 PM Pacific Daylight Time
From: Bloom PhD, Jesse D
To: Sergei Pond

Thanks Sergei. Hearing your words and critiques mean more to me than anything on that call, as I know how rigorous and honest you are.

From: Sergei Pond [redacted]
Sent: Sunday, June 20, 2021 3:37:08 PM
To: Bloom PhD, Jesse D [redacted]
Subject: Moral support

Dear Jesse,

Despite the guilt trip that Kristian was trying to put you on (not 100% sure why), I think what you are doing is the correct scientific approach. Report what you find and don't worry too much about political ramifications. Just wanted to offer my moral support as you face the inevitable shit-storm. Hopefully I was able to communicate some of my support on the call.

Best,
Sergei

E-mail 1: *My correspondence with Sergei Pond the afternoon immediately after the conclusion of the Sunday meeting with the NIH.*

Subject: Re: [External] URGENT
Date: Sunday, June 20, 2021 at 1:51:46 PM Pacific Daylight Time
From: Rasmus Nielsen
To: Bloom PhD, Jesse D

I do think you are doing the right thing. I have doubts too about many things - but I think it is right to bring these observations out. I just wish it wasn't so political.
Rasmus

On Jun 20, 2021, at 10:37 PM, Bloom PhD, Jesse D [redacted] wrote:

Thanks Rasmus. But do let me know honestly if you think I'm doing the wrong thing though. I have so much self doubt about this, I have you admit.

From: Rasmus Nielsen [redacted]
Sent: Sunday, June 20, 2021 1:29:39 PM
To: Bloom PhD, Jesse D [redacted]
Subject: Re: [External] URGENT

Hi Jesse,

That was a very difficult situation, but I think you handled it extremely well. Hope everything will be going well over the next few days. I think you are really standing up for science.

Best,
Rasmus

E-mail 2: *My correspondence with Rasmus Nielsen the afternoon immediately after the conclusion of the Sunday meeting with the NIH.*

Date: Monday, June 28, 2021 at 8:24:57 AM Pacific Daylight Time
From: Jon Cohen
To: Bloom PhD, Jesse D
CC: McElroy, Molly W

Thanks Jesse. I appreciate your transparent recounting. I'm not interested in gossip. If Fauci and Collins urged you not to publish, that would be an important point to clarify. But from you're recounting, that's not what happened.

I'll give the revised manuscript a read now.

Best,

Jon

On Jun 28, 2021, at 8:20 AM, Bloom PhD, Jesse D [redacted] wrote:

[EXTERNAL EMAIL]
Hi Jon,

Thanks for reaching out about this and the question about meeting with the NIH about the pre-print.

First, I wanted to let you know that I have revised the pre-print based on helpful feedback I received from a variety of scientists including Stephen Goldstein. Stephen kindly sent me his thoughts and posted them as a comment on *bioRxiv*, and I posted my response as a comment and revised the manuscript to address his comments and make a few other minor changes. All of this should show up on *bioRxiv* today or tomorrow.

I have attached the revised manuscript that I submitted to *bioRxiv*. One of the changes I made was to include the deletion e-mail request from Wuhan University as Figure 6 in this revised manuscript. I did not have this e-mail when I submitted the original version of the pre-print. But now I am getting so many questions about the reason for the deletion that I decided it would be most informative to just let people read the e-mail themselves rather than hear second-hand summaries. Although the wording is vague, some readings of the e-mail could imply that there were multiple deleted projects. It's important to figure out if that is the case and if so also analyze those other projects, just to be certain we have pursued all possible leads to learn as much as possible about early SARS-CoV-2.

As far as your question about meeting with the NIH:

After submitting the pre-print to *bioRxiv* on Friday (June 18), I e-mailed a copy to Francis Collins, Stephen Sherry (NCBI Director), and Toni Fauci to give them a heads up since this is a hot-button topic. They replied fairly quickly, including Stephen Sherry sending me the deletion request e-mail from Wuhan University that I have included as Figure 6 of the revised pre-print. They also suggested that we could meet on Sunday (June 20) to discuss the pre-print, and said they would invite Kristian Andersen and Bob Garry as outside experts. They asked if I thought anyone else should be invited, and I suggested Sergei Pond, Rasmus Nielsen, and Trevor Bedford, all of whom are experts on viral phylogenetics and have studied early SARS-CoV-2 sequences.

At that point, I e-mailed *bioRxiv* to ask them to hold the pre-print from posting until after the meeting on Sunday, out of courtesy to everyone else at the meeting. The meeting on Sunday included Stephen Sherry, Francis Collins, Toni Fauci, Kristian Andersen, Bob Garry, Sergei Pond, Rasmus Nielsen, and several others from the NIH. (Trevor Bedford was not able to attend). Stephen Sherry provided a nice summary of the SRA process for sequence deletions. There was then animated discussion of the pre-print, mostly involving Kristian Andersen and myself, and to a lesser degree Bob Garry, Rasmus Nielsen, and Sergei Pond. Kristian Andersen made clear that he strongly objected to my pre-print for various reasons, including that he thought it was unethical for me to analyze the data after the authors had asked for it to be deleted from the SRA. I made clear that despite Kristian Andersen's objections, I was going to move forward with posting the pre-print on *bioRxiv*. The NIH attendees said relatively little beyond asking some clarifying questions, and no one at the NIH asked me not to post the pre-print. Various people including Bob Garry and a few from the NIH made minor suggestions that I thought were constructive, but given the situation I decided it was better not to make any changes to the manuscript after the meeting. After the meeting concluded, I e-mailed *bioRxiv* again to ask them to move ahead with posting the pre-print after it had completed normal screening.

Because I was aware that the history of the pre-print could become a contentious topic, I placed the entire manuscript and analysis on GitHub where there are time-stamped commits tracking every change I have made through the whole history of the project (see here: https://github.com/jbloom/SARS-CoV-2_PRJNA612766/commits/main). As you can see, there are no changes to the manuscript between June 18 (when I e-mailed the manuscript to the NIH) and June 22 (when the initial version posted to *bioRxiv*). I did add one small additional analysis (commit ac5ce7a on June 19) to answer a good question that Francis Collins asked, which was whether it was possible that the sequences were in GISAID under another name (the answer is no, since the analysis shows that some mutation combinations are not present on GISAID).

More broadly, although I know talking about meeting with the NIH makes for juicy gossip, it is a distraction from what is actually the important question: are we doing everything possible to gather any additional scientific information that could be relevant for understanding the origins or early spread of SARS-CoV-2?

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

From: Jon Cohen [redacted]
Date: Monday, June 28, 2021 at 7:37 AM
To: Bloom PhD, Jesse D [redacted]
Subject: Fauci and Collins

Hi Jesse,

I'm revising the story for our print magazine and wanted to follow up about something I heard that may or may not be accurate.

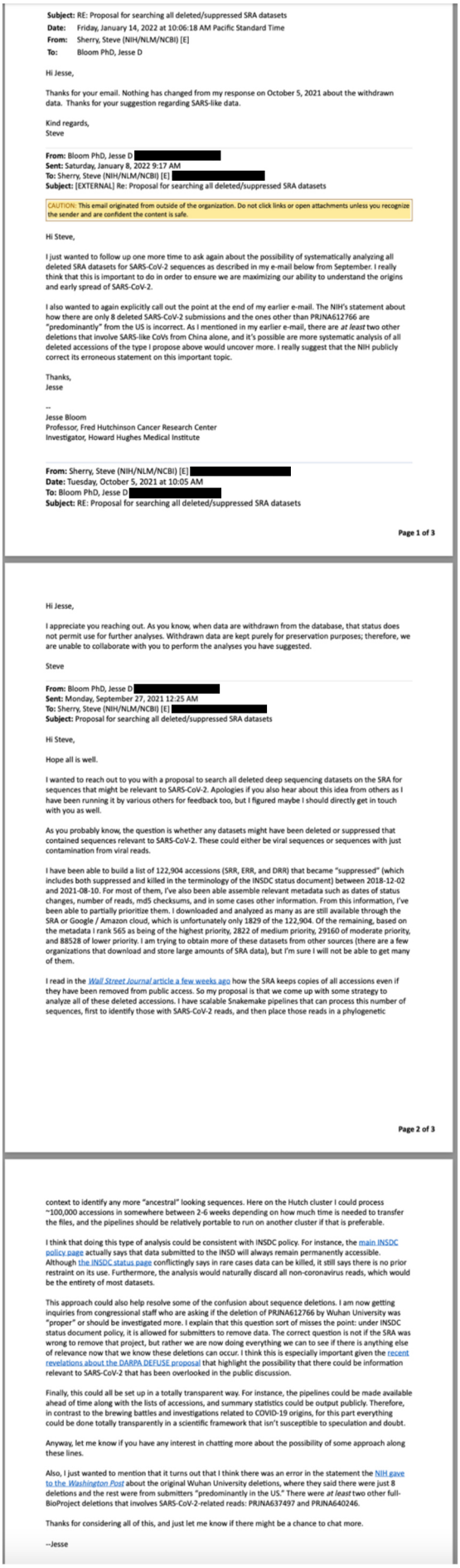
Did you have a discussion with Fauci and, possibly, Francis Collins, about publishing the paper? I was told they urged you not to publish.

If this did happen, can you describe any details? Was this a Zoom call? When? What did they say and what was your response?

Thanks,

Jon

Figure 3: *My correspondence with* Science *reporter Jon Cohen recounting the meeting one week later.*



Subject: RE: Proposal for searching all deleted/suppressed SRA datasets
Date: Friday, January 14, 2022 at 10:06:18 AM Pacific Standard Time
From: Sherry, Steve (NIH/NLM/NCBI) [E]
To: Bloom PhD, Jesse D

Hi Jesse,

Thanks for your email. Nothing has changed from my response on October 5, 2021 about the withdrawn data. Thanks for your suggestion regarding SARS-like data.

Kind regards,
Steve

From: Bloom PhD, Jesse D [redacted]
Sent: Saturday, January 8, 2022 9:17 AM
To: Sherry, Steve (NIH/NLM/NCBI) [E] [redacted]
Subject: [EXTERNAL] Re: Proposal for searching all deleted/suppressed SRA datasets

CAUTION: This email originated from outside of the organization. Do not click links or open attachments unless you recognize the sender and are confident the content is safe.

Hi Steve,

I just wanted to follow up one more time to ask again about the possibility of systematically analyzing all deleted SRA datasets for SARS-CoV-2 sequences as described in my e-mail below from September. I really think that this is important to do in order to ensure we are maximizing our ability to understand the origins and early spread of SARS-CoV-2.

I also wanted to again explicitly call out the point at the end of my earlier e-mail. The NIH's statement about how there are only 8 deleted SARS-CoV-2 submissions and the ones other than PRJNA612766 are "predominantly" from the US is incorrect. As I mentioned in my earlier e-mail, there are *at least* two other deletions that involve SARS-like CoVs from China alone, and it's possible are more systematic analysis of all deleted accessions of the type I propose above would uncover more. I really suggest that the NIH publicly correct its erroneous statement on this important topic.

Thanks,
Jesse

--
Jesse Bloom
Professor, Fred Hutchinson Cancer Research Center
Investigator, Howard Hughes Medical Institute

From: Sherry, Steve (NIH/NLM/NCBI) [E] [redacted]
Date: Tuesday, October 5, 2021 at 10:05 AM
To: Bloom PhD, Jesse D [redacted]
Subject: RE: Proposal for searching all deleted/suppressed SRA datasets

Page 1 of 3

Hi Jesse,

I appreciate you reaching out. As you know, when data are withdrawn from the database, that status does not permit use for further analyses. Withdrawn data are kept purely for preservation purposes; therefore, we are unable to collaborate with you to perform the analyses you have suggested.

Steve

From: Bloom PhD, Jesse D [redacted]
Sent: Monday, September 27, 2021 12:25 AM
To: Sherry, Steve (NIH/NLM/NCBI) [E] [redacted]
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.

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

Page 2 of 3

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

Figure 4: My more recent correspondence with NCBI Director Steve Sherry about deleted sequences of SARS-related CoVs.