

	<p>BioTeam, Inc.  <a href="http://www.bioteam.net">http://www.bioteam.net</a></p> <p><b>Primary Contact:</b>  Bhanu Rekepalli, Ph.D., Senior Director of Government Services  <b>Email:</b> bhanu@bioteam.net  <b>Tel:</b> 865-230-1605</p>
---	---

## NIH/NLM: Root Cause Analysis: Removal of SRA Sequence Data Records

Prepared for:

Kim Pruitt, Ph.D., Chief, Information Engineering Branch, NCBI  
Victor De La Torre, Service Area Manager/Division Director, CIT

Prepared by:

Ari Berman, Ph.D., CEO, BioTeam  
Laura Boykin, Ph.D., Senior Scientific Consultant, BioTeam  
Myra Ceasar, Senior Delivery Services Consultant, BioTeam  
Anna Sowa, Ph.D., Senior Scientific Consultant, BioTeam  
Simon Twigger, Ph.D. Director of Data Science, BioTeam

**Document History:**

**Draft Version 1, delivered August 6th, 2021**  
**Version 2, delivered August 12th, 2021**  
**Final Document, delivered August 20th, 2021**  
**Reformatted for Public Release, March 21st, 2022**

Executive Summary	4
1. Introduction	7
1.1. Introduction to the NLM SRA RCA Project	7
1.2. Root Cause Analysis Project Goals and Methodology	8
2. Defining the problem	10
2.1. Sequence status change terminology	11
3. Outline of Events	13
4. Facts and Documentation of the Situation	15
4.1. Detailed Timeline of Actions Taken on the 241 Wuhan SARS-CoV-2 Sequence Submission (BioProject PRJNA612766)	15
5. Identification of Possible Causal Factors	35
6. Problem Root Cause Evaluation	39
6.1. Step 1 - Submission	41
6.2. Step 2 - Submission troubleshooting	41
6.3. Step 3 - Processing & Value Add analyses	41
6.4. Step 4 - Standard Sequence replication	42
6.5. Step 5 - Ad hoc replication to COVID buckets	42
6.6. Step 6 - Status Change Request	43
6.7. Step 7 - Review and decision on the removal request	45
6.8. Step 8 - Action is taken based on the decision	50
Step 9 - Synchronization of SRA systems based on status change	52
6.9.	52
6.10. Step 10 - Monitoring of overall system consistency	53
6.11. Step 11 - Internal investigation into management of the sequences	54
6.12. Step 12 - Public response by NIH based on data gathered by internal investigation	56
7. Root Causes	59
7.1. Situation 1 - The public Wuhan sequences were killed rather than being suppressed	59
7.2. Situation 2 - Access to sequence files in the cloud is inconsistent with the current INSDC policies	60

7.3.	Situation 3 - 18 source data files have been lost, which does not meet NLM's goal of preserving its collected data	61
7.4.	Systemic Root Causes	62
8.	Future Considerations	64
8.1.	Evaluating potential opportunities	64
8.2.	Tier 1 Opportunities	66
8.2.1.	Policies and Procedures	67
8.2.2.	Communication	68
8.2.3.	IT Systems	69
8.3.	Tier 2 Opportunities	70
8.3.1.	Organization	70
8.3.2.	People	70
8.3.3.	Workload	71
8.3.4.	Environment	71
8.3.5.	Culture	71
9.	Conclusion	73
10.	Appendix	74
10.1.	NLM SRA RCA Interviewees (Anonymized)	74
10.2.	Scoring tables used in the FMEA	74
10.2.1.	Severity	74
10.2.2.	Occurrence	75
10.2.3.	Detection	75
10.3.	Documents List / Reference Materials	76

# Executive Summary

In June 2021, it became mainstream news that SARS-CoV-2 sequences sampled early in the pandemic (March 2020) from Wuhan, China were added to, then removed from the Sequence Read Archive (SRA) at the National Institutes of Health (NIH). The media gained knowledge of this event through a pre-print article by Dr. Jesse Bloom from the Fred Hutchinson Cancer Research Center on bioRxiv published on June 22, 2021 ([Bloom 2021a](#)). That preprint and the news articles alleged that since the origins of the COVID-19 pandemic have been called into question, these sequences might provide a critical insight into the early points of the virus's transmission in Wuhan Province. This series of events led NIH to launch internal investigations to understand what happened in this case.

The goal of this project was to independently assess, using a root cause analysis (RCA), the procedures, policies, and processes underlying the management of the 241 SARS-CoV-2 sequences deposited by a researcher from the School of Pharmaceutical Sciences at Wuhan University (Wuhan) on March 16th, 2020, to the Sequence Read Archive (SRA) at the National Institutes of Health (NIH), National Library of Medicine (NLM), National Center for Biotechnology Information (NCBI).

The specific problem statement that formed the focus of this analysis was defined as follows:

**“The sequences submitted by the group from Wuhan, China were not managed in a manner consistent with the current policies applicable to the NIH Sequence Read Archive.”**

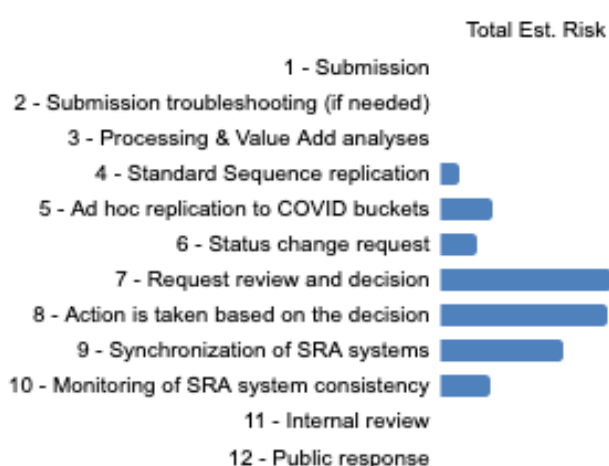
The RCA presented in this document clearly documents the sequence of events, the movement of the files in question, and the underlying issues that contributed to the problem that led to the publication of Dr. Bloom's preprint. This RCA solely focused on the sequences in question and general SRA issues that led to these events. ***Importantly, we find no obvious mal intent behind the actions taken during this incident on the part of SRA or its staff. We found a series of weaknesses in the system that led to this situation during an unprecedented public health emergency that has also been colored by political polarization on the topic of the COVID-19 pandemic origins.***

SRA was impacted by two main sets of policies, those of NLM to preserve data for the scientific record, and those of the International Nucleotide Sequence Database Collaboration (INSDC) which describe how sequence records will be managed by INSDC members like SRA. Three specific instances were discovered where management of the sequence records was not in line with the applicable policies:

1. The Wuhan sequences were public from March 16<sup>th</sup> 2020 until June 17<sup>th</sup> 2020 when, at the request of the original submitter, they were taken down from public access via the 'kill' command (intended for data that was never public), rather than via the 'suppress'

command which is intended for use on data that has been public, and simply made the sequences unsearchable in SRA but still accessible via their individual accession numbers, as intended by the INSDC policies.

2. Having been 'killed', the sequences in question should no longer be accessible via accession number within SRA. However, they are still present and accessible in various cloud-based locations operated by NCBI, highlighting a gap in current SRA procedures for handling sequences in the cloud that has led to the SRA system being in a state that is inconsistent with both its own and the current INSDC policies.
3. Examination of the detailed file movements and current file locations identified that, while processed data files exist for the entire Wuhan submission, 18 source sequence files were lost in various file movement operations out of the 241 submitted Wuhan sequences. This result does not meet NLM's goal of preserving its collected data.



The steps involved in the handling of the sequences were defined and analyzed in detail. These steps, along with a graph of the total estimated risk for each step, representing how much each step contributed to the problem under review are shown in the figure here.

A variety of root causes contributing to the incident were identified. Key amongst these were issues related to **Policies and Procedures, Communication (internal and external), and IT Systems**. Additional

'systemic' factors were also identified that were not direct contributors themselves, but which impacted many of the key factors – these included the **COVID 19 pandemic itself, the SRA budget, and the general priorities of SRA as a whole**.

### Future Considerations

Based on this analysis, BioTeam identified several potential opportunities that SRA could consider in order to mitigate the issues identified. These are presented in two tiers; Tier 1 focused on the largest issues that contributed to situation at hand. Highlights of the key opportunities identified in the Tier 1 opportunities include:

- Develop an appropriate cloud storage policy for SRA.
- Make improvements to internal policies and procedures.
- Acquire guidance from NIH on the role of SRA in public health emergencies and its data management responsibilities.
- Institute additional training for the curation team.
- Make improvements to the Microsoft Dynamics ticket tracking system used by the curators to handle communications with SRA users.

- Make improvements to the internal curation interfaces to improve efficiency and minimize the risk of inadvertent errors.

Tier 2 opportunities focus on areas to consider in future planning and in guiding improvements to the SRA system and operations. These include opportunities around workload, the organization and its culture, and ideas to support the SRA team. More details can be found in Section 8 of this document.

# 1. Introduction

The National Library of Medicine (NLM) is the world's largest biomedical library dedicated to making biomedical data and information more accessible. NLM enables researchers, clinicians, and the public to use the vast wealth of biomedical data to improve health.

The Sequence Read Archive (SRA), as part of NLM, is NIH's primary archive of high-throughput sequencing data. Their mission is to make sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. SRA is operated by the National Center for Biotechnology Information (NCBI) within NLM.

SRA accepts data from all sequencing platforms and as a member of International Nucleotide Sequence Database Collaboration (INSDC), shares data with partner INSDC members, which includes the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ).

As the volume of data in SRA has grown, the demand by the global community for the data it contains also grows. Starting during the 2020 SARS-CoV2 outbreak, SRA provides dedicated resources for SARS-CoV2 sequences and associated research. Currently, the NCBI SARS-CoV-2 resource database holds 929,176 SRA runs and is continuously growing.

## 1.1. Introduction to the NLM SRA RCA Project

This project centered around 241 SARS-CoV-2 sequences deposited by a researcher from the School of Pharmaceutical Sciences at Wuhan University (hereby referred to as Wuhan University) on March 16th, 2020, to SRA. The sequences were initially made public, and then subsequently withdrawn by request from that researcher on June 17th, 2020. In this document, any reference to Wuhan is restricted to these 241 SARS-CoV-2 sequences and the affiliated authors on the initial [Wang 2020b](#) publication. These Wuhan SARS-CoV-2 sequences were later described as “deleted” and “the trusting structures of science have been abused to obscure sequences relevant to the early spread of the SARS-CoV-2 in Wuhan” in a preprint by Dr. Jesse Bloom from the Fred Hutchinson Cancer Research Center on bioRxiv on June 22nd, 2021 ([Bloom 2021a](#)). Note: this paper was published in Molecular Biology and Evolution as of August 16<sup>th</sup>, 2021 ([Bloom 2021c](#)).

The goal of this project was to independently assess the procedures, policies, and processes underlying the withdrawal of the above-mentioned SRA records through a root cause analysis (RCA).

## 1.2. Root Cause Analysis Project Goals and Methodology

BioTeam collected much of the information that contributed to the root cause analysis through interviews of SRA staff and review of relevant documents provided by SRA. During the period of July 20th – 29th, 2021, BioTeam conducted 19 interviews with 18 members of NLM/SRA drawn from the SRA Operations and Developer Teams, plus representatives from NLM Leadership (see Table 1, below, for a descriptions of each group’s role). The interviews were intended to gain insight into any underlying organizational and structural issues surrounding the withdrawal of the 241 sequences from Wuhan University.

SRA group	Role
<b>SRA Operations Team</b>	The operations team consists of curators and curation team leads. This group works with the users of SRA to handle sequence submissions, updates, and other queries about finding and using data within SRA.
<b>SRA Developer Team</b>	The software development team builds and maintains much of the infrastructure and software systems used by SRA. This includes the on-premises compute and storage systems, the on-premises databases that store and track the data within SRA, some of the cloud storage buckets in AWS and Google Cloud platform, the SRA interfaces, and many of the various analysis pipelines that provide ‘value added’ analyses on top of the raw SRA datasets.
<b>Leadership</b>	In this context, Leadership refers to senior staff that are responsible for managing the various administrative units that SRA reports up to within NCBI.

*Table 1 - The names and roles of the three groups of NLM/SRA staff interviewed.*

Each SRA group was asked a series of questions to understand their role and the role of their team within the greater SRA landscape. They were also asked specifically if they had any involvement with the Wuhan SARS-CoV-2 sequences in question. The anonymized list of interviewees can be found at the end of this report.

BioTeam also referenced a variety of documentation during this analysis that were either directly provided to BioTeam by SRA or retrieved by BioTeam from the public web. These are summarized by category in Appendix 10.3 [Documents List / Reference Materials](#).

This report is the primary output of the above outlined work and draws on the interviews and research conducted specifically for this project as well as relevant research and documentation that are formally part of this project. **All findings and opportunities described in this document are exclusively relevant to the Wuhan sequences and their interactions with**



**SRA. A detailed analysis of the entirety of SRA was not performed as a part of this project.**

This document and the underlying analysis were guided by a core set of **principles**:

- The focus is on the root cause and not simply the symptoms of the problem
- There can be, and often are, multiple root causes
- The focus is on the how and the why something happened, not on the who is responsible
- A goal was to be methodical and find concrete evidence to back up root cause claims

Thus, while analyzing the SRA platform as it relates to the Wuhan sequences, we took a comprehensive and holistic look at the entire ecosystem. In addition to discovering the root cause, BioTeam strives to provide context and information that can result in meaningful action by outlining potential opportunities.

## 2. Defining the problem

A Root Cause Analysis begins with the definition of the situation under consideration. For the purposes of this root cause analysis, our scope was as follows:

### **When did the situation occur?**

The time period under consideration for this analysis begins on March 16th, 2020 when the sequences were first submitted to SRA, and ends on Thursday, July 9th, 2021 when BioTeam was officially able to begin a third-party review of the situation.

### **Where did it occur?**

The sequences were uploaded to SRA and processed on systems housed at NIH and then released to the public. As part of the release process, the sequences were also copied onto several commercial cloud environments used by SRA for the storage and dissemination of its data collection.

### **What was the impact?**

Following a request from the submitter for SRA to retract/withdraw the sequences from the SRA database, action was taken by SRA to meet this request and the sequences in question were no longer indexed in the SRA system and not searchable in the system. The sequences were also not accessible in any way from the main on-premises SRA systems. However, some of these sequences were still able to be accessed in specific cloud storage locations. This raised two primary questions:

- Was the request for the sequences' removal handled correctly by SRA?
- Why were the corresponding sequence files inaccessible via SRA but still accessible via the cloud?
- Are all the associated files accounted for in the SRA ecosystem?

### **What is the core problem?**

As an NIH sequence database and member of the International Nucleotide Sequence Database Consortium (INSDC), SRA operates according to a variety of policies and procedures defined by this structure. Similarly, as part of NLM, SRA is subject to NLM's policies around collection and preservation<sup>1</sup>. These policies can be considered as threshold criteria against which SRA's actions can be evaluated to answer the three questions above.

As an example, INSDC policies<sup>2</sup> state the following:

---

<sup>1</sup> The Collection and Preservation Policy of the NLM, <https://www.ncbi.nlm.nih.gov/books/NBK518808/> last accessed, 8/5/21

<sup>2</sup> INSDC Policy statement, <https://www.insdc.org/policy.html> last accessed, 8/4/21

*“3. All database records submitted to the INSD 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”*

In this situation, the records submitted to SRA were made public, a request was made to remove these records, however, these records were then unavailable via accession from the main SRA but still available via accession in the cloud.

Based on this, **the core problem is that the sequences submitted by the group from Wuhan, China were not managed in a manner consistent with the current policies applicable to the NIH Sequence Read Archive**. The following root cause analysis explores the causes of this situation and seeks to identify potential opportunities that can prevent similar situations happening in the future.

## 2.1. Sequence status change terminology

There are various words used to describe the status of data at the SRA and INSDC partners (see Figure 1 1, below) and an understanding of these phrases will be useful to fully understand the material that follows. In the case of this set of sequences (Figure 1 1 part A, below), the Wuhan researcher wrote to SRA requesting retraction (Comms Documents) of the sequences and a BioProject curator wrote back and suggested that the researcher use the option “replaced by”. A few days later an SRA curator used the word “withdraw” and the action of “killing” the sequences was invoked. There was confusion around the terminology being used to change the status of these sequences and what actions should be taken to address the submitter’s request.

The official INSDC status options are shown in green with the corresponding terms used in SRA policies as of June 2020 shown to the left (Policy Documents). To the right of the INSDC terms are shown the terms used in SRA policies as of June 2021 (Policy Documents). A new term has appeared (‘live’), the term ‘redact’ is no longer used to refer to the INSDC term ‘killed’, however, the policies now use the INSDC term ‘killed’ and document a new term (‘withdrawn’) as two ways to achieve the same end result (the affected data is not indexed in SRA and is not accessible via a direct link using the sequence’s accession number).

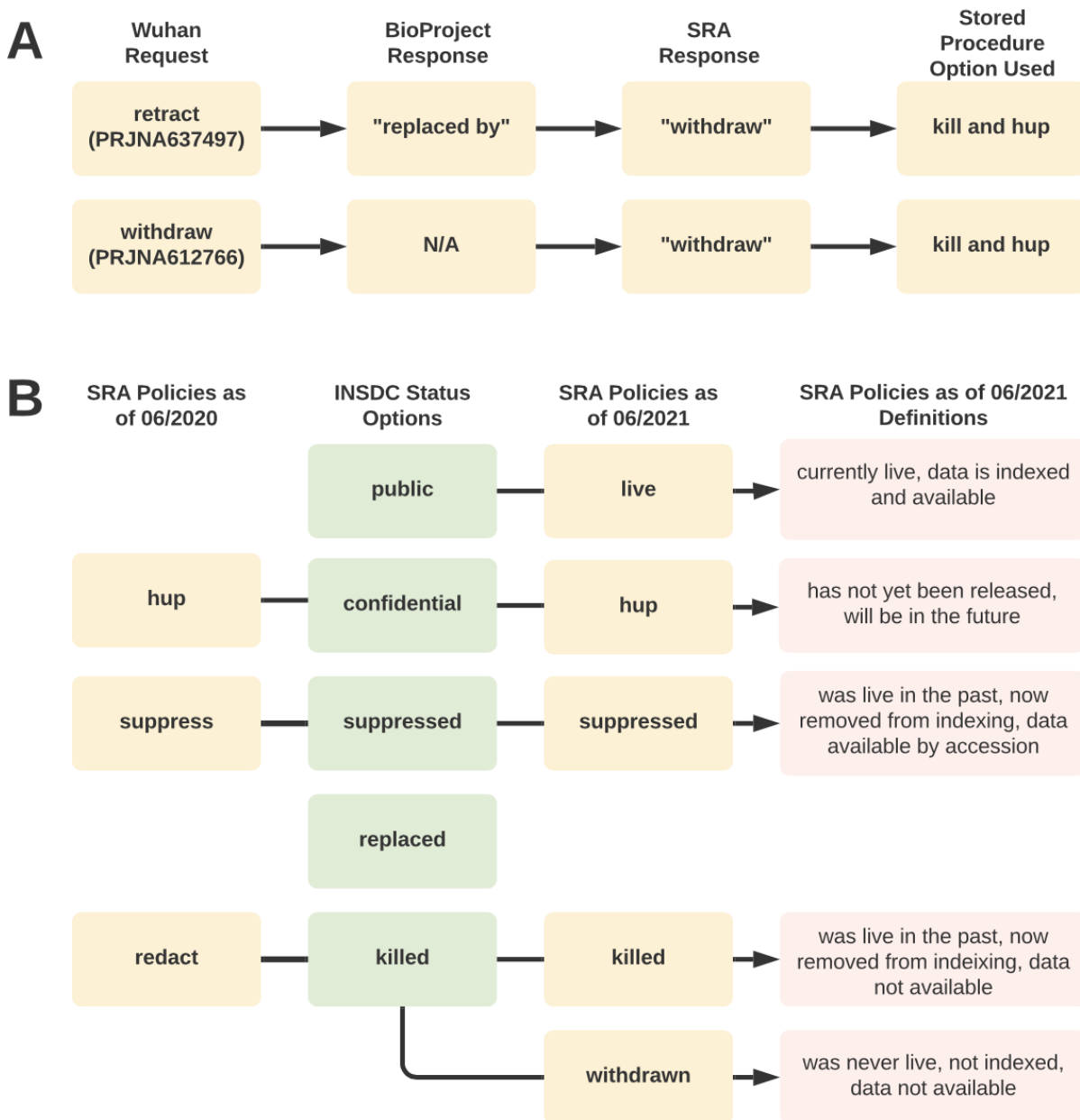


Figure 1 1 – Part A: The terminology that was used in the communications surrounding the Wuhan SARS-CoV-2 sequences in 2020 and B) the terminology and definitions of relevant terms and their relationship to INSDC language (green boxes). Part B illustrates the potential for confusion around sequence status nomenclature.

### 3. Outline of Events

The timeline in Figure 2 2 below outlines the publications and key events in the submission, release, and subsequent withdrawal of all files associated with BioProject PRJNA612766.

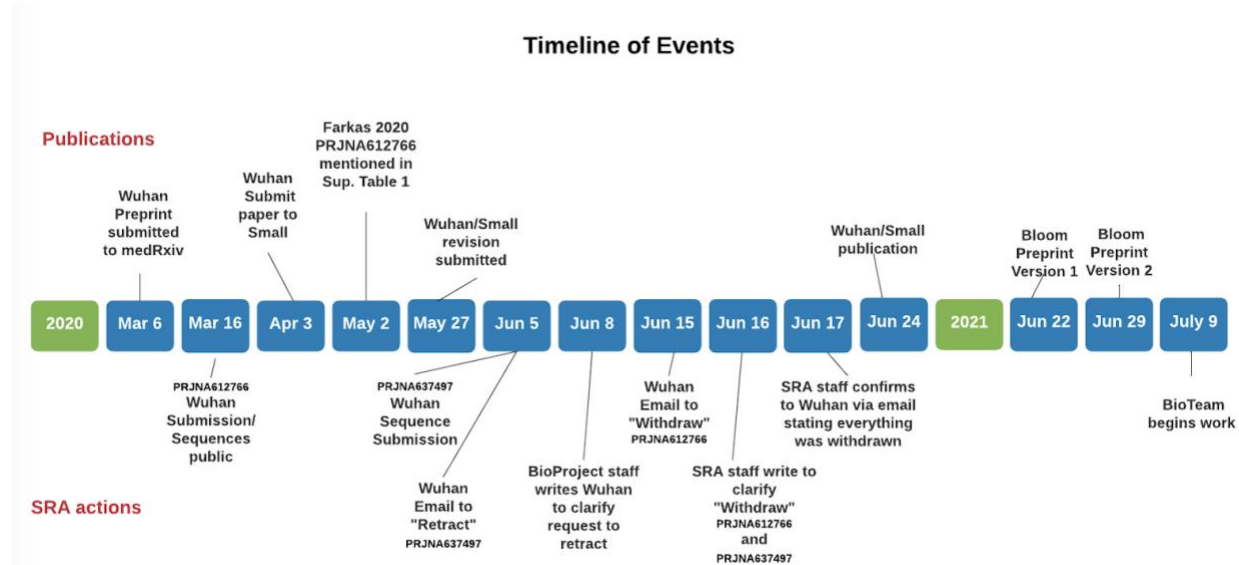


Figure 2 2 - Timeline of the major events surrounding the submission, release and subsequent withdrawal of all files associated with BioProject PRJNA612766

- 1) These data were first mentioned in a preprint (Wang 2020a) posted to medRxiv on March 6th, 2020 and the main goal of this study was to show the utility of Oxford Nanopore Technologies in detecting SARS-CoV-2 in patient samples taken at various time points (241 amplicons generated). SRA was not made aware of these data until the submission via the SRA submission site 10 days after the preprint was online.
- 2) Submission of the sequences was successful, the associated reference numbers (SRR11313269-SRR11313509) PRJNA612766 were emailed to the submitter on March 16, 2020 (Comms Documents), and the sequences were made public (HisPSV).
- 3) The Wuhan team subsequently submitted their publication to the journal Small (Wang 2020b) on April 3rd, 2020.
- 4) While Wang (2020b) was in review at Small, Farkas (2020) obtained all available SARS-CoV2 sequences in SRA and downloaded the data. The purpose of the Farkas 2020 paper was to describe early mutational events across samples from publicly available SARS-CoV-2 sequences. Wang (2020b) also mentions a revision of the initial submission was received by the editorial staff on May 27th, 2020.
- 5) The same submitter from the Wang publication (2020a and 2020b) submitted another BioProject PRJNA637497 to SRA that contained 1 sequence on June 5th, 2020 (Comms Documents).
- 6) The submitter of PRJNA637497 and PRJNA612766 requested that the data be "retracted" 14 hours after the submission of PRJNA637497 (Comms Documents).

- 7) On June 8th, 2020 a member of the BioProject curation staff wrote back to the submitter asking for clarity and also pointing out editing the submission is preferred to deleting the submission. An additional 2 emails were sent by SRA curation staff verifying the submitter has two submissions with SRA. Continuing that email conversation the submitter writes back on June 15th, 2020, to request both submissions PRJNA637497 and PRJNA612766 to be withdrawn (Comms Documents).
- 8) The final email correspondence for the incident was on June 17th, 2020, and the SRA curator confirmed everything was withdrawn (Comms Documents).
- 9) Wang 2020b was published on June 24th, 2020, with no mention of the SRA accession numbers.
- 10) Bloom 2021a outlines while researching the early spread of SARS-CoV-2 he went to obtain all the data listed in Farkas (2020) and found BioProject PRJNA612766 unavailable via a google search and also “no items found” from the NCBI SRA search box. This led Bloom to start looking for the sequences in the cloud (he noticed cloud links to other SRA submissions) and he found them located in a google cloud bucket <https://storage.googleapis.com/nih-sequence-read-archive/run/<ACCESSION>/<ACCESSION>>.
- 11) Bloom 2021b was available online 7 days after version1 with updates including a redacted version of the SRA email responses to the submitter, inclusion of 2 runs (SRR11313490 and SRR11313499) he recovered from archives downloaded before June 2020, adjustment to the trimming of the reads and several other wording clarifications.
- 12) July 5th, 2021 sequences from Wang (2020b) were submitted to the [Genome Sequence Archive \(GSA\)](#) hosted by the China National Center for Bioinformation (CNCB) and made available on July 8th, 2021.
- 13) July 9, 2021, BioTeam began the RCA.

## 4. Facts and Documentation of the Situation

A critical component of this root cause analysis was to trace the sequence of events concerning the Wuhan SARS-CoV-2 sequences in question in as much detail as possible taken from as many sources as possible into a single timeline. In this section, we outline our detailed findings of all known actions that were taken regarding the Small 2020a submission, leading up to the “kill” procedure initiated in response to the user’s request, and following those events up through the current date. The goal of this section is to provide the reader with a clear view of the facts and actions that led to the need to perform this RCA.

We present the sequence of events in a timeline format describing the following information:

- Date
- Time (if known)
- Action taken
- The person or system that took the action
- The result of the action
- The implications
- Any errors or examples taken from the source data to illustrate the action

The actions timeline is limited only to those events that directly concern the Wuhan SARS-CoV-2 sequences and artifacts and any major SRA-level events that may have shaped the resulting actions or consequences of those actions.

Most of the information in this events timeline was taken from detailed log files that document all recorded actions taken on the Wuhan SARS-CoV-2 sequences from the point of submission in March 2020, through the end of July 2021. These log files (Log Files) were provided to us by SRA staff by combining information from all parts of the SRA database, joined into a single file relating to this submission only. The log file (Log Files) was 15,919 lines long including details of every action taken on every file throughout its lifecycle within the SRA system. This log information has been condensed and summarized in the timeline below to be human readable and to give enough information to clearly explain each event and its implications.

### 4.1. Detailed Timeline of Actions Taken on the 241 Wuhan SARS-CoV-2 Sequence Submission (BioProject PRJNA612766)

As demonstrated in Figure 3 3 below, the 241 SARS-CoV-2 sequences in question were submitted to SRA on March 16th, 2020. Upon submission, the runs were not marked with a publication date and were set to immediately be made public after the processing of the submission was completed by the system. These sequences and their artifact files generated as



a part of the SRA submission process went through a very large number of file movements across SRA systems and cloud locations throughout their lifecycle.

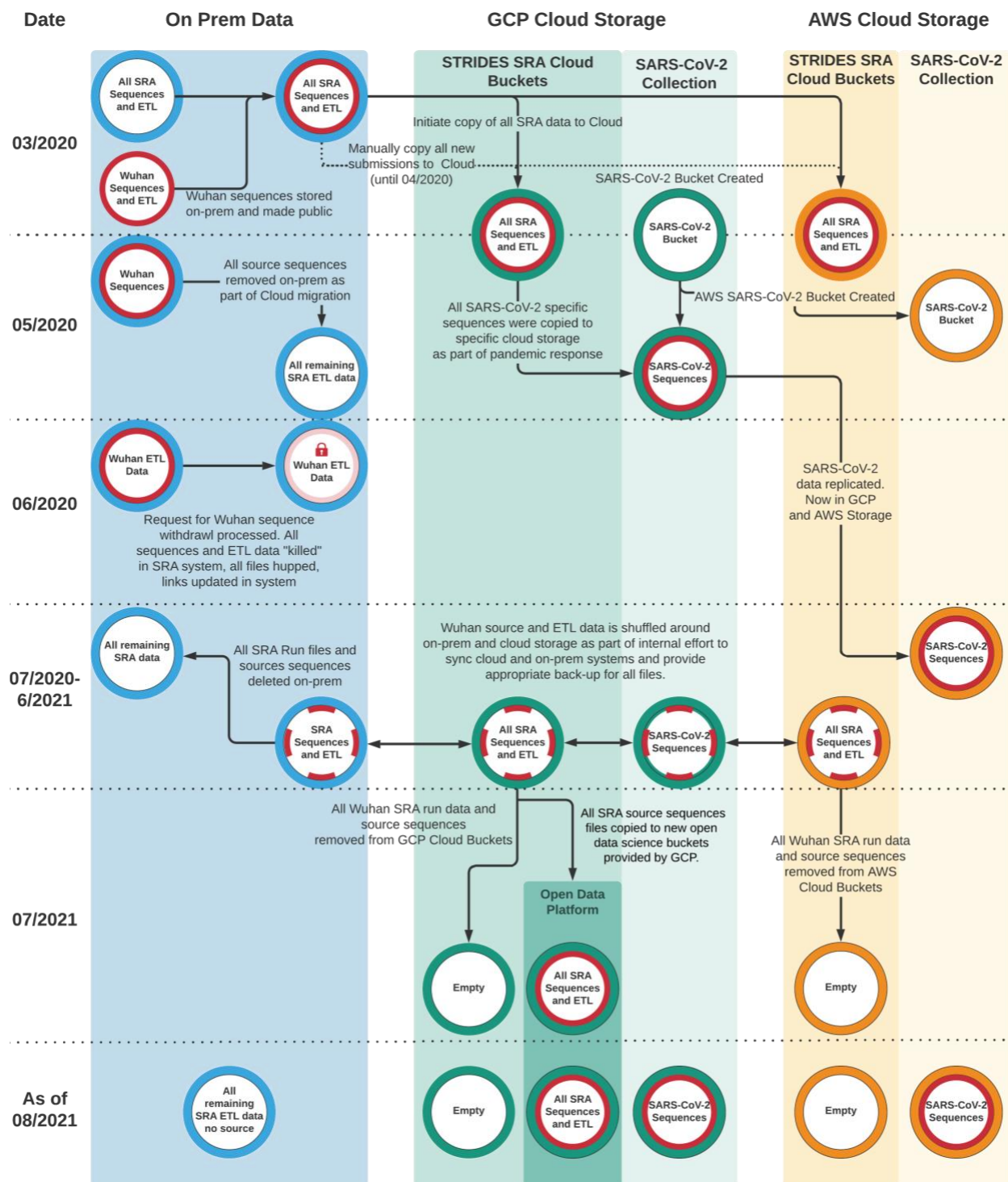


Figure 3 3 – SARS-CoV-2 Sequences Submission / Movement, and Timeline of Actions Taken.



As is evident in Figure 3 3, the sequences in question went through a lot of location changes, copies, updates, and movement throughout the lifespan of their residence in the SRA system. Many of these movements were due to policy changes or SRA-level initiatives that were being implemented in parallel, like the movement of SRA to the cloud to mitigate constrained storage space within NCBI's datacenters. In fact, in March through April 2020, movement of data to the cloud was an imperative action to keep SRA functional as it was running out of on-premises space very quickly. Many manual copy processes were initiated that weren't all error-checked or always tracked. Due to some failures in the data transfers, some of the sequences from the Wuhan group were lost in the process. In total, 18 source files (.fastq) were lost from PRJNA612766, these are listed below and were confirmed by staff member L4<sup>3</sup>.

SRR11313269	SRR11313339	SRR11313398	SRR11313425	SRR11313440	SRR11313484
SRR11313280	SRR11313364	SRR11313410	SRR11313431	SRR11313476	SRR11313490
SRR11313300	SRR11313371	SRR11313421	SRR11313438	SRR11313477	SRR11313491

Table 2 - 18 source files (.fastq) from BioProject PRJNA612766 that have been confirmed as missing.

The detailed summary of actions within the SRA system taken on the Wuhan SARS-CoV-2 sequence data are listed in the following dated entries. SRA-level activities or actions that had a direct impact on the fate of the sequences in question are also listed on the timeline and are marked in **BLUE**. Actions taken specifically on the Wuhan SARS-CoV-2 sequences are not colored, except for the execution of the "kill" request by staff member O2<sup>4</sup>, which is highlighted in **RED**.

3/2020

*Initiated Copy of all SRA data to the Cloud*

**Action:** SRA local storage out of space, initiated copy of all data to GCP and AWS  
**Taken By:** Developer Team  
**Result:** SRA sequences and artifacts now available in the cloud  
**Implication:** Secondary location - more to keep track of, 45PB of data, 13M runs, 30M sequences  
**Notes/Errors:** 5% data loss from failed data transfers and subsequent data deletion to save space

3/2020

*Manual copy of all new submissions to cloud*

<sup>3</sup> L4 is an anonymized designation for one of the interviewees in the Leadership group. A complete list of the anonymized interviewees is available in Appendix 10.1

<sup>4</sup> O2 is an anonymized designation for one of the interviewees in the Operations group. A complete list of the anonymized interviewees is available in Appendix 10.1

**Action:** All new submissions manually copied to cloud  
**Taken By:** Developer Team  
**Result:** Manual copy of sequences available in cloud  
**Implication:** Manual processes are prone to error and lack of consistency checking with the core SRA system. This is how the Wuhan SARS-CoV-2 sequences got to the cloud  
**Notes/Errors:**

3/16/2020 1:38 PM

*Wuhan Sequences Submitted to SRA*

---

**Action:** 241 Nanopore SARS-CoV-2 Amplicon Sequences from Wuhan patients submitted to SRA  
**Taken By:** Wuhan University School of Pharmaceutical Sciences  
**Result:** Uploaded to SRA  
**Implication:** Sequences were marked public upon submission  
**Notes/Errors:**

3/16/2020 3:44 PM

*Wuhan Sequences Submitted to SRA*

---

**Action:** Files loaded into SRA and Stats calculated  
**Taken By:** SRA System  
**Result:** Ready for release  
**Implication:** Sequences ready for release  
**Notes/Errors:** SRR11313290 - warning: Spot '3031a1c3-85dd-4547-950c-ee552d399a60' has already been assigned a spot id warning: Spot '10f2933f-77aa-4e64-99b1-24b25ac4375d' has already been assigned a spot id warning: Spot '9a779cba-84ca-44a9-82a7-815d0732051d' has already been assigned a spot id warning: Spot '4e52879a-bd06-4856-86ac-3e81f6e035ed' has already been assigned a spot id warning: Spot '5f13fb1b-b7a0-45c0-953c-0a7d8df98715' has already been assigned a spot id warning: Spot '44c1f6d9-a6ae-41d0-95cb-780ac6f24a28' has already been assigned a spot id warning: Spot 'e0e131a7-c320-4407-829e-3b4217304dfb' has already been assigned a spot id warning: Spot '41aa7a7c-d7e3-40a0-b61d-29c6d87b7c98' has already been assigned a spot id warning: Spot 'd3a43b1f-083d-4915-a9ff-778ef96f6cec' has already been assigned a spot id warning: Spot '596db866-6fb8-4985-bc1d-e16bcd78bd2f' has already been assigned a spot id

3/16/2020 7:29 PM

*Wuhan SRA Runs Released to Public*

---

**Action:** All files processed and released on-prem  
**Taken By:** SRA System  
**Result:** All sequences and stats public and available  
**Implication:**  
**Notes/Errors:**

3/16/2020 11:45 PM

*Wuhan SRA Runs Copied to the Cloud*

---

**Action:** SRA run copied to GCP and AWS hot buckets:  
s3://sra-pub-run-4/SRR11313403/SRR11313403.1  
gs://sra-pub-run-2/SRR11313403/SRR11313403.1  
gs://sra-pub-run-8/SRR11313499/SRR11313499.1  
**Taken By:** Developer Team  
**Result:** Wuhan SRA Run now publicly available in GCP and AWS  
**Implication:** Since data was publicly released on 3/16/20 and cloud migration was underway, a public cloud copy of the data was ok  
**Notes/Errors:**

3/17/2020 6:44 AM

*Taxonomy Analysis Completed*

---

**Action:** Taxonomy Analysis completed on the sequences  
**Taken By:** Curation Team  
**Result:** Verification that data did not contain human data, verification that sequences were SARS-CoV-2  
**Implication:** Data was verified to be correctly described by the submitter  
**Notes/Errors:**

3/17/2020 11:09 PM

*More SRA Runs Copied to Cloud*

---

**Action:** More Wuhan SRA runs copied to GCP and AWS hot buckets:  
gs://sra-pub-run-8/SRR11313290/SRR11313290.1

s3://sra-pub-run-9/SRR11313490/SRR11313490.1  
s3://sra-pub-run-9/SRR11313290/SRR11313290.1  
gs://sra-pub-run-8/SRR11313490/SRR11313490.1  
s3://sra-pub-run-9/SRR11313499/SRR11313499.1  
gs://sra-pub-run-8/SRR11313288/SRR11313288.1  
s3://sra-pub-run-9/SRR11313288/SRR11313288.1

**Taken By:** Developer Team

**Result:** Wuhan SRA Run now publicly available in GCP and AWS

**Implication:** Since data was publicly released on 3/16/20 and cloud migration was underway, a public cloud copy of the data was ok

**Notes/Errors:**

---

4/2020

*SARS Detection Automated*

**Action:** SARS detection pipeline and copy into COVID bucket automated

**Taken By:** Developer Team

**Result:** Manual detection by Curation Team no longer required

**Implication:**

**Notes/Errors:**

---

4/2020

*Copy to Cloud Automated*

**Action:** Copy of submitted sequences to cloud automated

**Taken By:** Developer Team

**Result:** Manual copy of newly submitted data no longer required

**Implication:**

**Notes/Errors:**

---

4/8/2020 3:27 AM

*Second Taxonomy Analysis Performed*

**Action:** Another Taxonomy Analysis was completed on the sequences

**Taken By:** SRA System

**Result:** Revalidation that sequences are correct, realignment with deeper SARS-

CoV-2 reference

**Implication:** Unclear why the tax analysis was redone

**Notes/Errors:**

5/12/2020 12:55 PM

*Wuhan Source Sequence Files Copied to Cloud*

**Action:** All 241 Wuhan source sequences copied to Google hot storage bucket  
gs://sra-pub-src-8/

**Taken By:** Developer Team

**Result:** Wuhan source sequences now publicly available in Google cloud

**Implication:** Since data was publicly released on 3/16/20 and cloud migration was underway, a public cloud copy of the source sequences was ok

**Notes/Errors:**

5/12/2020 1:26 PM

*Wuhan Source Sequence Files Deleted On-Premises*

**Action:** Wuhan sequences were deleted from on-prem storage

**Taken By:** Developer Team

**Result:** Wuhan sequences now only available from the Google cloud bucket

**Implication:** Local source sequences were deleted from on-prem storage from all of SRA to save storage space that was running out quickly. The Wuhan sequences were included in that decision.

**Notes/Errors:**

5/26/2020

*SARS-CoV-2 Specific Collection Created in AWS Cloud*

**Action:** SARS-CoV-2 specific collection was created on AWS S3 in response to the pandemic

**Taken By:** Developer Team

**Result:**

**Implication:**

**Notes/Errors:**

5/27/2020 6:53 AM

*Wuhan Data Copy to SARS-CoV-2 Collection in Cloud Failed*

---

**Action:** Initiated copy of Wuhan data to AWS S3 SARS-CoV-2 Bucket  
**Taken By:** Developer Team  
**Result:** Copy of ETL artifacts failed due to error, could not be completed  
**Implication:** Example error  
**Notes/Errors:** adjust\_location\_copy failed :: ... failed with rc=1; copy failed: s3://sra-pub-run-4/SRR11313283/SRR11313283.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313283/SRR11313283 An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

5/28/2020 7:54 AM

*Wuhan Source Sequences Copied to GCP SARS2 Collection*

---

**Action:** Wuhan source sequences were copied from GCP Hot bucket to GCP SARS-CoV-2 Bucket  
copy from gs://sra-pub-src-4 to gs://sra-pub-sars-cov2 done  
**Taken By:** Developer Team  
**Result:**  
**Implication:**  
**Notes/Errors:**

5/29/2020 12:08 AM

*Wuhan SRA Data Removed from GCP SARS-CoV-2 Collection*

---

**Action:** SRA run information and ETL data removed from GCP SARS-CoV-2 specific bucket:  
removed from gs://sra-pub-sars-cov2  
**Taken By:** Developer Team  
**Result:** Only the source sequence data remains in the SARS-CoV-2 bucket, rather than all of the artifact data submitted or created through the ETL process.  
**Implication:** The SARS-CoV-2 bucket was intended to contain only source sequence information. It is likely that the other SRA information was copied there unintentionally.  
**Notes/Errors:**

5/30/2020 9:06 AM

*Wuhan SRA Data Removed from AWS SARS-CoV-2 Collection*

---

- Action:** SRA run information and ETL data removed from AWS SARS-CoV-2 specific bucket:  
removed from s3://sra-pub-sars-cov2
- Taken By:** Developer Team
- Result:** Only the source sequence data remains in the SARS-CoV-2 bucket, rather than all of the artifact data submitted or created through the ETL process.
- Implication:** The SARS-CoV-2 bucket was intended to contain only source sequence information. It is likely that the other SRA information was copied there unintentionally.
- Notes/Errors:** The logs showed that the initial copy of this information to the S3 SARS-CoV-2 bucket failed, with no indication that the copy had been re-initiated. It is presumed that the copy failed

6/2/2020 3:56 PM

*Failed (2): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

---

- Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error
- Taken By:** Developer Team
- Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error
- Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example error:
- Notes/Errors:** adjust\_location\_copy failed :: ...ailed with rc=1; copy failed: s3://sra-pub-src-8/SRR11313487/R14-4h.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313487/R14-4h.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/3/2020 3:58 PM

*Failed (3): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

---

- Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error

**Taken By:** Developer Team

**Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error

**Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example error:

**Notes/Errors:** adjust\_location\_copy failed :: ...ailed with rc=1; copy failed: s3://sra-pub-src-8/SRR11313283/D10-4h.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313283/D10-4h.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/4/2020 3:58 PM

*Failed (4): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

---

**Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error

**Taken By:** Developer Team

**Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error

**Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example error:

**Notes/Errors:** adjust\_location\_copy failed :: ...ailed with rc=1; copy failed: s3://sra-pub-src-8/SRR11313283/D10-4h.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313283/D10-4h.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/5/2020 9:08 AM

*SRA Data Updated in SRA Database*

---

**Action:** Some of the SRA ETL and experiment artifacts release date was updated in the SRA database:  
SRX8476671  
SRR11931188  
tmp\_75.fastq  
SRS6776624

**Taken By:** Developer Team

**Result:** Release date, execution of Taxonomy analysis, ETL and links were recreated for these artifacts

**Implication:** This was presumably done to update the information as more reference



information became available on the submission. It could also have been done to kick the system into re-initiating a new automated sync to the cloud

**Notes/Errors:**

6/5/2020 3:59 PM

*Failed (5): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

---

**Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error

**Taken By:** Developer Team

**Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error

**Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example Error:

**Notes/Errors:** adjust\_location\_copy failed :: ...d with rc=1; copy failed: s3://sra-pub-src-4/SRR11313286/C2-10min.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313286/C2-10min.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/5/2020 6:38 PM

*SRA Data Updated in SRA Database*

---

**Action:** Updated ETL and experiment data for Wuhan sequences were copied to GCP SRA Hot Bucket:  
copied to gs://sra-pub-run-3/SRR11931188/SRR11931188.1  
copied to gs://sra-pub-src-8/SRR11931188/tmp\_75.fastq.1

**Taken By:** Developer Team

**Result:** Updated information now available in GCP buckets

**Implication:**

**Notes/Errors:**

6/6/2020 4:00 PM

*Failed (6): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

---

**Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error

**Taken By:** Developer Team

**Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error

**Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example Error:

**Notes/Errors:** adjust\_location\_copy failed :: ...d with rc=1; copy failed: s3://sra-pub-src-4/SRR11313286/C2-10min.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313286/C2-10min.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/7/2020 4:02 PM

*Failed (7): Wuhan Source Sequences Copied to AWS SARS2 Bucket*

**Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, again with an error

**Taken By:** Developer Team

**Result:** No source sequences were copied into the S3 bucket due to an AccessDenied error

**Implication:** Sequences do not exist in the AWS SARS-CoV-2 archive. Example Error:

**Notes/Errors:** adjust\_location\_copy failed :: ...d with rc=1; copy failed: s3://sra-pub-src-4/SRR11313286/C2-10min.fastq.1 to s3://sra-pub-sars-cov2/sra-src/SRR11313286/C2-10min.fastq An error occurred (AccessDenied) when calling the CopyObject operation: Access Denied

6/17/2020 1:01 PM

*Wuhan Data Killed in SRA*

**Action:** All Sequence and ETL data were killed in the SRA system, all files hupped, links updated in the database

**Taken By:** Interviewee [O2](#)

**Result:** Wuhan data could no longer be searched for or accessed through the SRA interface and the files were put into a non-accessible state (hold until published)

**Implication:** Sequences and artifacts from the Wuhan dataset were no longer available to the public from the on-prem systems, but they were not deleted from the system. The source sequences weren't present on-prem because they were moved to GCP on 5/12/2020 in order to save room

on on-prem systems

**Notes/Errors:** Detailed logs of the results of issuing the kill procedure are documented in HisPSV. Note that the data was NOT deleted by the “kill” procedure, merely marked HUP.

7/23/2020 1:00 PM

*Wuhan Source Sequences Copied to AWS SARS2 Bucket*

**Action:** Wuhan source sequences were copied again to the AWS S3 SARS-CoV-2 specific bucket, this time successfully:  
copy from s3://sra-pub-src-4 to s3://sra-pub-sars-cov2 done

**Taken By:** Developer Team

**Result:** Source sequence files were successfully copied to the AWS S3 SARS-CoV-2 bucket

**Implication:** Source sequences now publicly available in both GCP and AWS SARS-CoV-2 specific collections

**Notes/Errors:**

7/29/2020 4:12 PM

*More Wuhan Source Sequences Copied to AWS SARS2 Bucket*

**Action:** Selected sequences from the Wuhan set were copied to the AWS S3 SARS-CoV-2 bucket. These source files were not previously copied to the S3 bucket:  
respiratory-10min.fastq  
respiratory-2h.fastq  
0cp-replicate04-1h.fastq  
R09-10min.fastq

**Taken By:** Developer Team

**Result:** Missing source sequence files from the Wuhan submission were added to the S3 SARS-CoV-2 Bucket

**Implication:** These sequences might have been from the second, smaller submission by the Wuhan group in March, and were missed during the first copy attempts.

**Notes/Errors:**

7/31/2020 4:33 AM

*Selected Wuhan Source Files Copied On-Premises*

**Action:** Some of the Wuhan source files were retrieved from S3 and moved to on-premises storage:

D2-10min.fastq  
3000cp-replicate01-1h.fastq  
3000cp-replicate03-4h.fastq  
F5-10min.fastq  
E5-4h.fastq  
E1-4h.fastq  
3000cp-replicate04-4h.fastq  
3000cp-replicate04-10min.fastq  
3000cp-replicate04-2h.fastq  
3000cp-replicate03-10min.fastq  
A1-4h.fastq  
D12-10min.fastq  
3000cp-replicate03-30min.fastq  
C1-10min.fastq  
C11-10min.fastq  
3000cp-replicate01-10min.fastq  
3000cp-replicate03-1h.fastq  
3000cp-replicate02-4h.fastq  
3000cp-replicate02-1h.fastq  
1000cp-replicate03-4h.fastq  
3000cp-replicate02-30min.fastq  
1000cp-replicate02-30min.fastq  
1000cp-replicate02-10min.fastq  
1000cp-replicate01-1h.fastq  
F12-4h.fastq  
C2-4h.fastq  
3000cp-replicate04-1h.fastq  
3000cp-replicate02-10min.fastq  
A2-10min.fastq  
3000cp-replicate02-2h.fastq  
E5-10min.fastq  
1000cp-replicate04-10min.fastq  
3000cp-replicate03-2h.fastq  
1000cp-replicate02-1h.fastq  
1000cp-replicate04-2h.fastq  
1000cp-replicate03-1h.fastq

**Taken By:** Developer Team

**Result:** Source files were again placed on on-prem storage from this dataset

**Implication:** The sequences were originally lost from the on-prem storage systems due to transfer errors from backup to the cloud. Perhaps these were copied back to be added back to the backups.

**Notes/Errors:**

10/24/2020 9:57 AM

*All Wuhan Data Removed from SRA Cloud Buckets*

---

**Action:** All 241 Wuhan source sequences were removed from the following AWS and GCP buckets:  
removed from s3://sra-pub-src-8  
removed from s3://sra-pub-src-4  
removed from gs://sra-pub-src-4

**Taken By:** Developer Team

**Result:** Source sequences were no longer publicly available in the SRA cloud Hot Buckets

**Implication:** We assume that because the state sync with the SRA database never occurred when the Wuhan dataset was killed on 6/17/2020.

**Notes/Errors:**

11/27/2020 3:49 PM

*Wuhan SRA Data Files Copied to AWS SARS2 Bucket*

---

**Action:** 239 of the Wuhan SRR files were copied to the AWS S3 SARS-CoV-2 Bucket:  
s3://sra-pub-sars-cov2

**Taken By:** Developer Team

**Result:** SRA artifacts were now available through the SARS-CoV-2 S3 Bucket, where they had been removed on 5/29/2020 and 5/30/2020

**Implication:** These files were originally removed from the buckets because they weren't source sequence files but SRA artifacts. The decision must have been made to add them back to the archive at some point

**Notes/Errors:**

11/27/2020 3:56 PM

*Some Wuhan Source Sequences Copied to GCP SRA Bucket*

---

**Action:** 12 of the Wuhan source sequence files were copied from on-prem

storage to the SRA GCP Hot Bucket again:  
gs://sra-pub-src-4/

**Taken By:** Developer Team

**Result:** The 12 source files were now publicly accessible from the GCP bucket after being removed on 10/24/2020

**Implication:** It is unclear why the sequences were removed, and then only 12 put back

**Notes/Errors:**

2/23/2021 5:25 PM

*All Wuhan Source Data Entries in SRA Database were Updated*

---

**Action:** All 241 Wuhan SRS entries in the SRA database were smp updated, modified, marked released, and re-linked in SRA

**Taken By:** Developer Team

**Result:** The Wuhan SRA run was once again available for search and public in the SRA database and interfaces

**Implication:** It is unclear why the sequences were re-released

**Notes/Errors:** Note - even though the logs show the sequences were re-released, they remain unsearchable in the SRA interface

2/23/2021 5:25 PM

*Selected Wuhan Source Data Killed Again*

---

**Action:** 51 of the Wuhan SRS entries in the SRA database were immediately killed again, leaving 190 still public in the database

**Taken By:** Developer Team

**Result:** 190 of the 241 original SRS files in the dataset remained publicly available after the last two operations.

**Implication:** It is unclear why the 51 sequences were killed, and not the other 190. It appears that this release was a mistake and not entirely reversed

**Notes/Errors:**

6/16/2021 10:20 PM

*All Wuhan SRA Data Files Removed from GCP SARS2 Bucket*

---

**Action:** All 241 SRA Run files were removed from the GCP SARS-CoV-2 specific

bucket: gs://sra-pub-sars-cov2

**Taken By:** Developer Team

**Result:** The SRA ETL and quality data were no longer available in the GCP SARS-CoV-2 Bucket

**Implication:** We assume this was done so that only the original sequence files remained in the archive

**Notes/Errors:**

6/16/2021 10:26 PM

*Selected Wuhan Source Files Removed from GCP SARS2 Bucket*

---

**Action:** 18 of the Wuhan source sequence files were removed from the GCP SARS-CoV-2 specific bucket.

**Taken By:** Developer Team

**Result:** These 18 source sequence files were no longer publicly available in the GCP SARS2 Bucket

**Implication:** It is unclear why these specific sequences were removed from the archive

**Notes/Errors:**

6/18/2021 12:27 AM

*All Wuhan SRA Runs and Source Files Deleted On-Premises*

---

**Action:** All Wuhan SRA Run files and source sequence files were deleted from on-prem storage

**Taken By:** Developer Team

**Result:** These files no longer existed on the on-prem storage systems for SRA

**Implication:** Unclear why this action was taken

**Notes/Errors:**

6/15/2021

*All SRA Data Copied to ODP Buckets on GCP*

---

**Action:** All SRA files were copied to the new open data platform buckets provided by GCP (free storage for SRA):

gs://nih-sequence-read-archive

**Taken By:** Developer Team

**Result:** All SRA files were now publicly available from the ODP bucket in GCP

**Implication:** This change was made to save money for SRA. Many SRA run files and source sequence files were moved to the ODP buckets because they were being provided for free, and removed from on-prem and paid-for cloud buckets.

**Notes/Errors:**

6/24/2021 7:07 PM

*All Wuhan SRA Run Files Copied to ODP Buckets on GCP*

---

**Action:** All Wuhan SRR files were copied to the new open data platform buckets provided by GCP (free storage for SRA):  
gs://nih-sequence-read-archive

**Taken By:** Developer Team

**Result:** The Wuhan SRA run files were now publicly available from the ODS bucket in GCP

**Implication:** This change was made to save money for SRA. Many SRA run files and source sequence files were moved to the ODS buckets because they were being provided for free and removed from on-prem and paid-for cloud buckets.

These runs were never authorized to be re-released by SRA, it is unclear why they were, what the decision chain on that decision was, and why the researcher was never notified

**Notes/Errors:**

6/28/2021 5:26 PM

*All Wuhan Source Files Copied to ODP Buckets on GCP*

---

**Action:** All Wuhan source sequence files were copied to the new open data platform buckets provided by GCP (free storage for SRA):

gs://nih-sequence-read-archive

**Taken By:** Developer Team

**Result:** The Wuhan source sequence files were now publicly available from the ODS bucket in GCP



**Implication:** Same comment as the previous entry

**Notes/Errors:**

7/2/2021

*All SRA Data was Removed from SRA Cloud Buckets*

**Action:** All SRA run data and source sequence files were removed from the GCP and AWS STRIDES Hot Buckets

**Taken By:** Developer Team

**Result:** The SRA data was now publicly available from the ODS bucket in GCP and not from any other SRA storage location (except SARS2 archives)

**Implication:** Same comment as the previous entry

**Notes/Errors:**

7/3/2021 12:11 AM

*All Wuhan Data was Removed from SRA Cloud Buckets*

**Action:** All Wuhan SRA run data and source sequence files were removed from the GCP and AWS Hot Buckets

**Taken By:** Developer Team

**Result:** The Wuhan data was now publicly available from the ODS bucket in GCP and not from any other SRA storage location (except SARS2 archives)

**Implication:** Same comment as the previous entry

**Notes/Errors:**

7/6/2021 8:23 PM

*SRA Database Updated to Reflect New Wuhan Location in ODS*

**Action:** All SRA database links and statuses for the Wuhan submission were updated to reflect the new location in the ODS bucket on GCP.

**Taken By:** Developer Team

**Result:** The Wuhan dataset information in the database was updated to reflect the new location in the ODS GCP bucket

**Implication:**

**Notes/Errors:**

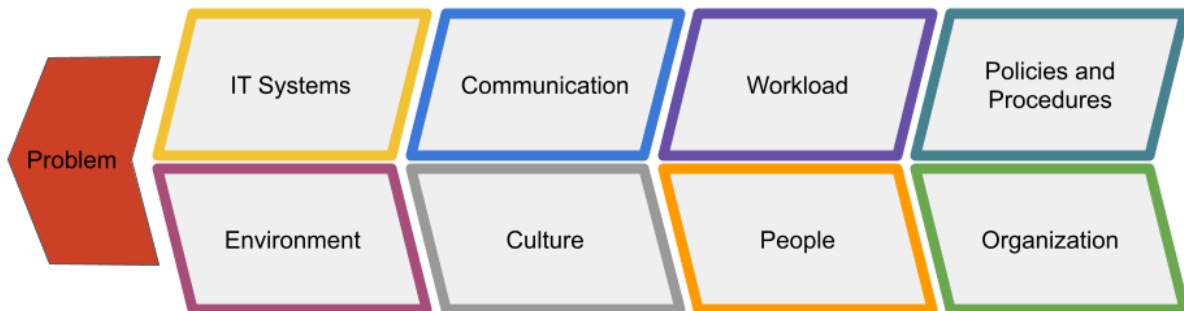


## 5. Identification of Possible Causal Factors

The first step in identifying the root cause of a problem is the identification of possible causal factors that exist within the framework of the problem.

A fishbone diagram is an RCA tool that can be used to structure a brainstorming session and to sort ideas and causes into categories.

The goal of this section is to understand the conditions that allowed the problem to occur and what other issues co-exist with the central problem. BioTeam identified eight categories relevant to the root cause of the problem as identified in Figure 4 below.



*Figure 4 4 - Fishbone diagram representing the higher-order categories of potential causes leading to the identified problem for BioProject PRJNA612766.*

**IT Systems** refers to the compute and technical infrastructure that are utilized by SRA (both on-premises and in the cloud). These systems are created and maintained by SRA's developers and used by SRA's operations team to run the processes with SRA. Identified potential causal factors related to IT systems are:

1. Cloud: The SRA has transitioned to the cloud as part of an internal re-prioritization effort. Movement to cloud environments brings with it a number of challenges that include meeting the expectations of cloud providers, training of internal staff, adapted policies for cloud storage, modification of production systems for cloud use, and the development of new user-facing guidance and documentation.
2. Dynamics ticketing system: the dynamics ticketing system is used for tracking tickets and communications. The system in 2020 was not adequately adapted to the needs of SRA staff and SRA workflows. Dynamics is a barrier to both operations staff as well as leadership who cannot easily review tickets.
3. Curation interfaces and legacy systems: legacy systems and many current interfaces are developed using multiple scripting languages. The maintenance of these legacy systems is difficult coupled with the need for interfaces to provide easier workflows and UI/UX makes these systems challenging. Connected to this is the lack of documentation and training around these systems and interfaces.

**Communication** refers to the ability to transfer relevant information consistently and effectively across the organization. Identified potential causal factors related to communication are:

1. To align SRA staff to guiding policies, it is necessary to communicate the definition of the specific words related to the stored procedures and data categories to include the definition of “withdraw”, “kill” or “suppress” both internally and externally to the users of SRA.
2. The SRA supports a diverse user base and is thus in need of very clear and specific user-facing documentation to guide users and manage user expectations.
3. SRA needs clear internal documentation and procedures that can easily be communicated to new and current staff.
4. The flow of internal communications proceeds in a point-to-point manner where individuals in part of the organization prefer to communicate to another specific individual within the organization to resolve issues. Point-to-point communications preclude the dissemination of information and can lead to siloed workflows.
5. Managing responses to public and government inquiries requires an effective internal flow of information and organizational structure to support information gathering.
6. SRA strategic plans should be shared across the organization.

**Workload** refers to two specific aspects: the total amount of work that is present relative to the number of staff as well as to the way that the work is prioritized. Workload also considers the nature of the work, the complexity of the work, the focus on individual vs teamwork, and resource availability. Identified potential causal factors related to workload are:

1. The operations team manages a large influx of tickets daily. Therefore, automation, easy processing interfaces, and documentation can be used to reduce their workload.
2. The organizational priorities are focused on moving to the cloud, making COVID-19 data more accessible, launching new features, and increasing customer ease of access to data. These organizational priorities have stretched developer and support capacity.
3. The SRA team has a high technical debt of tasks that are undone and need to be prioritized for long-term internal success.

**Policies and Procedures** refers to the internal, organizational, government, and external policies and procedures that exist which are meant to guide the work and mission. Current relevant policies and procedures are to include:

- The Collection and Preservation Policy of the NLM
- INSDC policies
- SRA Success Metrics
- Federal Records Policy
- Internal standard operating procedures (SOPs)

Identified potential causal factors related to policies and procedures are:

1. Written documentation is necessary to align activities within SRA. Lacking internal documentation for procedures and SOPs negatively impacted the diverse organization.
2. Due to the transition of SRA to the cloud, policies and procedures need to be adjusted to align current internal and external policies to fit with the capabilities and attributes of data in the cloud.
3. Although overarching policies exist, such as those described above, they are often vague and difficult to translate to actionable direction.
4. The policies and procedures for how to proceed in cases of high priority or national security are either not available or are not well understood.

**Environment** refers to the totality of external factors that were happening during the time of the incident that contributed directly or indirectly to the problem. Identified potential causal factors related to the environment are:

1. During the period of 2019 and 2020, there was ongoing pressure to move SRA data to the cloud.
2. The necessary adjustments that were made to accommodate COVID-19 pandemic shifted the way that work was done on a day-to-day basis as well as the needs of researchers to access and make sense of data quickly.
3. Budgetary constraints not only put pressure on a database that is continually growing in size, but also the specific funding constraints of the time period in question added additional confusion.

**People** refer to the human factors such as staff, diversity, training, and expectations that contributed to the problem. People also consider aspects related to individual knowledge and capabilities, group norms, knowledge and skills, experience, and motivation/attitude. Identified potential causal factors related to people are:

1. The operations and development staff often have unique skill sets and lack skill redundancy, which means that a large portion of knowledge is siloed or limited to one individual.
2. Additional training for operations and developers is needed in order to tackle new demands of the SRA platform.
3. User expectations are not adequately managed with either public-facing documentation or user agreements.
4. The diversity of users causes language barriers and/or miscommunications
5. The lack of diversity on the SRA team could be limiting the SRA team's ability to effectively communicate with the end users.

**Culture** refers to the diversity, demographics, cultural competency, internal understanding of incentives, transparency, voluntary error reporting, information sharing, and a willingness to follow the internal set of standards. Identified potential causal factors related to culture are:

1. Within the SRA, there is a culture of being reactive (versus proactive) to the growing list of technical tasks that need to be accomplished and updated. With the move to the cloud, that list has greatly expanded and has further stretched the staff.
2. Overall, there is a culture of caring within the organization; everyone cares about doing the right thing, doing things well, and enabling science.
3. There are two, sometimes opposing, cultures and missions at the SRA - one of being an academic archive and one of supporting public health.
4. The strategic plan of the SRA highlights the need to meet metrics of being in the cloud and having the data easily accessible. There are many other tasks that are not prioritized officially but rather prioritized internally like quick turnaround times with tickets.
5. The desired culture appears to be one of innovation.

**Organization** refers to the structures that support SRA. Organization refers to the way the teams are organized, how checks and balances are established, the political climate, economic pressures, government incentives, new technology, leadership involvement, and group norms. Identified potential causal factors related to the organization are:

1. There are no established checks and balances or QA/QC steps for operations.
2. There is a lack of oversight into SRA by NLM as a whole.
3. There is no knowledge redundancy within the organization.
4. SRA appears to be siloed from other NCBI databases.
5. The organizational structure is not clear, and it is difficult to gauge who should know what subset of information.

The above-identified factors are considered potential causal factors, the specifically identified root causes connected to each step of the process are identified in Section 6 below.

## 6. Problem Root Cause Evaluation

The section above provides a high-level view of the main categories of factors identified that contributed to the problem being investigated. This section describes a more detailed analysis of the specific steps that took place related to the management of the sequences in question, and the initial activities that took place in response to the notification to NIH that there were questions about the status of these sequence records within SRA.

This section is structured as a Failure Modes and Effects Analysis (FMEA), which is a step-by-step approach for identifying all possible failures for this RCA. It has been adjusted to reflect only contributing failure modes. Failure modes are errors or defects, especially those that affect operations. As part of an FMEA analysis, failures are then scored and prioritized according to how serious their consequences are, how frequently they might occur, and how easily they can be detected. The purpose of an FMEA is to take actions to eliminate or reduce the highest priority areas. An FMEA starts off with understanding the steps of a particular process. The high-level steps involved are shown in Figure 5 Figure 5 5 - Outline of the 10 Process Steps directly considered in the FMEA analysis, plus the two post-incident actions that were reviewed but which did not contribute to the incident itself. below, along with a general summary of what is happening at each point in the process.

As part of the detailed analysis, each of these steps was evaluated in turn to identify the various 'inputs' to that process step and, for those inputs which were determined to have contributed to the problem being evaluated, a further analysis was performed to outline the following information:

- In what ways did the process input fail?
- What was the impact of this failure, relevant to the problem under consideration?
- One a scale of 1-10, what is the estimated Severity (SEV) of this failure?
- What caused the failure?
- One a scale of 1-10, what is an estimate of how often this failure might occur (OCC)?
- What controls currently exist, if any, that could have detected or prevented the failure?
- One a scale of 1-10, what is the estimated likelihood that current controls could detect (DET) or prevent the observed failure?

For inputs determined to have contributed to the problem at hand, an estimate of overall impact of that input's failure was generated by the simple formula:  $SEV * OCC * DET = \text{Risk Priority Number (RPN)}$ . The scoring schemes used to estimate Severity, Occurrence, and Detection are provided in the Appendix. The full detailed analysis will be submitted as a part of the supporting documentation for reference.



Figure 5 5 - Outline of the 10 Process Steps directly considered in the FMEA analysis, plus the two post-incident actions that were reviewed but which did not contribute to the incident itself.



To provide a high-level overview, Figure 66 below shows the total combined estimated risk score for each of the above steps. This score was calculated by combining the risk scores of each contributing issue in each step to come up with a total combined score for each step as a whole.

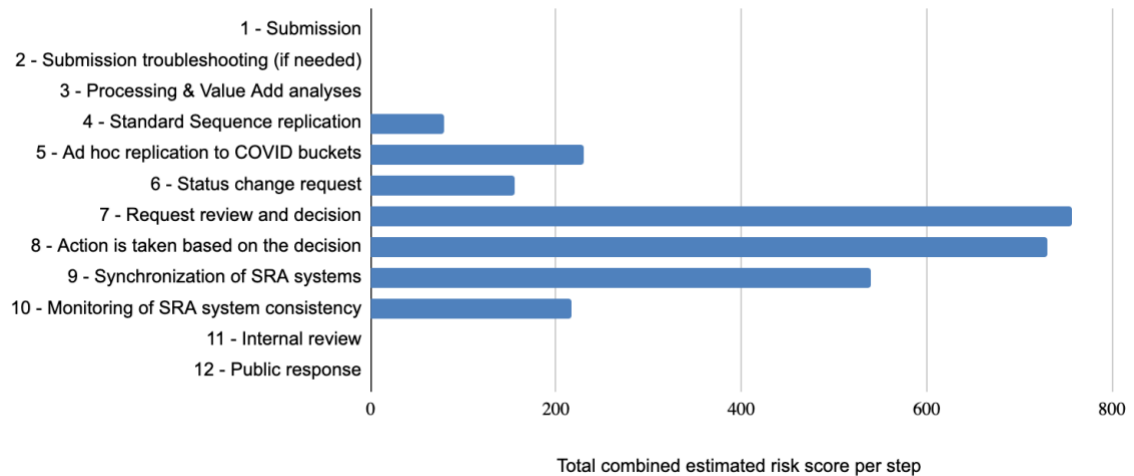


Figure 66 - Graph of total estimated risk score for each of the identified process steps.

From this chart, it is clear that the main areas of concern focused on how the request for the removal of the sequences was assessed (Step 7); how the removal process was enacted in the SRA system (Step 8); and how the associated changes propagated across SRA as a whole. The section below summarizes the key areas of risk within each step.

## 6.1. Step 1 - Submission

There were no factors contributing to the problem identified in this step.

## 6.2. Step 2 - Submission troubleshooting

There were no issues loading the sequence submission from Wuhan so this step was not needed.

## 6.3. Step 3 - Processing & Value Add analyses

There were no factors contributing to the problem identified in this step.

## 6.4. Step 4 - Standard Sequence replication

This step consisted of a variety of individual events responsible for replicating the relevant sequences across the SRA system and into the normal cloud buckets. These steps performed as expected.

However, in the same time frame that these sequences were being processed, SRA's on-premises storage was running critically low on available space, leading to the decision to replicate the entire SRA database from the standard on-premises backup storage locations, into GCP cloud storage. During the course of this migration 18 out of 241 source data files were lost.

### Internal Procedures

*As part of the Ad hoc replication to COVID buckets*

<b>Input:</b>	<b>Scripts for replicating files to the cloud</b>
<b>Failure:</b>	Incomplete backup of source data
<b>Impact:</b>	Movement of sequences copied to the COVID buckets were not initially tracked until a later time.
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• Failure to replicate some on-prem files to the cloud due to undetected errors during the urgent movement of on-prem archives to the cloud</li></ul>
<b>Current controls:</b>	<ul style="list-style-type: none"><li>• Review of log files combined with analysis of sequence files existing in various locations to identify any missing files</li></ul>
<b>Estimated Risk:</b>	<b>80</b> => 8 SEV * 2 OCC * 5 DET

## 6.5. Step 5 - Ad hoc replication to COVID buckets

### Internal Procedures

*As part of the Ad hoc replication to COVID buckets*

<b>Input:</b>	<b>Internal procedures</b> for moving data to newly created/temporary/high priority buckets
<b>Failure:</b>	The system did not keep track of the sequences as they were copied to a new location
<b>Impact:</b>	Movement of sequences copied to the COVID buckets were not initially

	tracked until a later time.
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>• No SOPs existed for managing and tracking sequences moved to non-standard buckets</li> <li>• Urgent nature of the situation may have caused documentation generation to have been deemphasized</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>• There are no current controls for this situation.</li> </ul>
<b>Estimated Risk:</b>	<b>140</b> => 7 SEV * 2 OCC * 10 DET

## 6.6. Step 6 - Status Change Request

### SRA Website

*As part of the Status Change Request*

<b>Input:</b>	<b>SRA Website</b>
<b>Failure:</b>	Options for removing sequence records from SRA were not effectively explained
<b>Impact:</b>	<ul style="list-style-type: none"> <li>• Submitter was not directed to the appropriate course of action prior to emailing SRA staff</li> <li>• Submitter had no prior understanding of what was possible, based on the appropriate INSDC policies</li> </ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>• Perhaps no problems had previously been identified that could be traced to poor documentation</li> <li>• Workload and focus on processing submissions had precluded regular updates to the public-facing website</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>• Limited FAQ information - <a href="https://www.ncbi.nlm.nih.gov/sra/docs/submitquestions/#question3upd">https://www.ncbi.nlm.nih.gov/sra/docs/submitquestions/#question3upd</a></li> <li>• Links to the INSDC status code documentation as the only other explanatory material: <a href="https://www.insdc.org/documents/insdc-status-document">https://www.insdc.org/documents/insdc-status-document</a></li> </ul>
<b>Estimated Risk:</b>	<b>36</b> => 2 SEV * 6 OCC * 3 DET

## INSDC Website

*As part of the Status Change Request*

<b>Input:</b>	<b>INSDC Website</b>
<b>Failure:</b>	Defines the INSDC status codes but does not indicate how/if sequence statuses can transition from one status to another
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Submitter may not know what options are available due to unclear explanation of what a currently public sequence record can transition to following a request to 'remove' those sequences from an INSDC database</li><li>• Submitter may not know what the implications of a status change are to the visibility and accessibility of their sequences due to unclear depiction of the practical implementations of these different status codes at an INSDC database</li></ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• INSDC website is focused on basic documentation and may not be envisaged as a primary source of documentation for sequence submitters (perhaps expecting this would be expanded upon by the INSDC member websites directly)</li></ul>
<b>Current controls:</b>	<ul style="list-style-type: none"><li>• Overarching INSDC Policy statement: <a href="https://www.insdc.org/policy.html">https://www.insdc.org/policy.html</a></li><li>• List of INSDC status codes: <a href="https://www.insdc.org/documents/insdc-status-document">https://www.insdc.org/documents/insdc-status-document</a></li></ul>
<b>Estimated Risk:</b>	24 => 2 SEV * 3 OCC * 4 DET

## BioProject Website

*As part of the Status Change Request*

<b>Input:</b>	<b>BioProject Website</b>
<b>Failure:</b>	Contains no references to removing BioProject entries on the BioProject website, nor any discussions of the implications for child entries (e.g. associated SRA sequence records)
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Submitter was not directed to the appropriate course of action prior to emailing SRA staff</li><li>• Submitter had no prior understanding of what was possible, based on the appropriate INSDC policies</li></ul>

<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>Online documentation focuses on the standard submission processes (which do reference other NIH dbs) but provide little to no coverage of the exceptions, such as when deleting from BioProject, and what the implications are for related database entries such as SRA</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>NLM has a Head of Sequence Archives position that oversees SRA, GenBank and other relevant NCBI databases. This position may provide some oversight on the integration of the various NCBI sequence archives.</li> </ul>
<b>Estimated Risk:</b>	46 => 2 SEV * 3 OCC * 8 DET

#### Information from submission process

*As part of the Status Change Request*

<b>Input:</b>	<b>Documents and information obtained during submission process</b>
<b>Failure:</b>	Documentation received following successful submission contained a contact email to use in the case of updates, it did not explain policies and procedures for any subsequent actions such as HUP, Suppress, etc.
<b>Impact:</b>	<ul style="list-style-type: none"> <li>Submitter was not directed to the appropriate course of action prior to emailing SRA staff</li> <li>Submitter may have had unrealistic expectations of what was possible, based on the appropriate INSDC policies</li> </ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>Providing additional detailed information on updates and withdrawal procedures, etc. may not have been considered</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>There is an email address to contact if the submitter wishes to update their record</li> </ul>
<b>Estimated Risk:</b>	48 => 3 SEV * 2 OCC * 8 DET

## 6.7. Step 7 - Review and decision on the removal request

## Email correspondence

*As part of the Review and decision on the removal request*

<b>Input:</b>	<b>Submitters email and subsequent communications with Submitter</b>
<b>Failure:</b>	Submitter first requested to remove the relevant BioProject entry, the BioProject curator suggested updating instead but said that it could be deleted if appropriate, no mention made of any associated child objects. When asking about the SRA submissions, standard email templates were not used to suggest other approaches, or to inform the submitter of the consequences of this request to remove records from SRA.
<b>Impact:</b>	<ul style="list-style-type: none"><li>• The submitter persisted in requesting removal of the sequences from SRA (rather than keeping it public or the lesser option of suppression)</li></ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• Lack of written SOPs that cover the appropriate ways to respond to submitters in various scenarios</li><li>• Curator was not aware of standard templates</li><li>• No central, authoritative, source containing all available documentation (docs appear to be found in at least two locations)</li><li>• Language barrier between submitter and SRA staff</li><li>• MS Dynamics is not well configured to support the reuse of standard templates</li><li>• Lack of effective ongoing training and feedback to ensure all staff are in sync around communication with submitters</li></ul>
<b>Current controls:</b>	<ul style="list-style-type: none"><li>• Existing 'discourage_suppress_request_form_letter' is available</li><li>• Initial curator training</li><li>• Weekly curation team meetings</li><li>• Ad hoc internal curation team discussions</li></ul>
<b>Estimated Risk:</b>	<b>90</b> => 2 SEV * 5 OCC * 9 DET

## INSDC policy

*As part of the Review and decision on the removal request*

<b>Input:</b>	<b>INSDC policy</b>
<b>Failure:</b>	SRA internal language for status changes diverges from INSDC policy language
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Potential confusion over implementation of SRA commands and how</li></ul>

	they relate to INSDC policies
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>• Legacy systems and commands developed prior to adoption of INSDC policies and terms</li> <li>• Lack of effective ongoing training and feedback to ensure all staff are in sync around the implementation of INSDC policies at SRA</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>• Initial curator training</li> <li>• Weekly curation team meetings</li> </ul>
<b>Estimated Risk:</b>	<b>120</b> => 5 SEV * 3 OCC * 8 DET

## Internal policies

*As part of the Review and decision on the removal request*

<b>Input:</b>	<b>Internal policies</b>
<b>Failure:</b>	<ul style="list-style-type: none"> <li>• Did not specify use of email templates</li> <li>• Did not adequately discuss status change options</li> <li>• Did not adequately relate SRA status change options to overarching INSDC policies</li> <li>• Did not adequately define official escalation process (what to do) or threshold conditions (when to use) that should be applied in situations such as this (request from submitter, failing system commands)</li> <li>• Do not have provisions for extraordinary situations (such as a global pandemic) that may require specific modifications to standard policies and procedures</li> </ul>
<b>Impact:</b>	<ul style="list-style-type: none"> <li>• Official templates were not used potentially leading to ineffective communication with the submitter</li> <li>• A decision was made by the curator without the opportunity for input from senior leadership</li> <li>• An incorrect action decision was reached that did not reflect the appropriate INSDC policies that should be applied in this situation</li> <li>• Potentially high-profile data was processed without a clear documentation chain that could be easily used in a retrospective review to determine what happened and why.</li> </ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>• Internal documentation is missing relevant content</li> <li>• Internal documentation is not reviewed and updated on a regular basis</li> </ul>

- Reliance on unwritten rules and shared knowledge and assumptions that this knowledge is universally known, correctly understood, and reflects the correct approach to be taken
  - Relevant policies and procedures around handling questions/exceptional circumstances have not been defined
  - Lack of effective ongoing training and feedback to ensure all staff are in sync around the implementation of INSDC policies at SRA
  - No formal processes for identifying extraordinary situations (such as a global pandemic) and enacting modified procedures to better respond to these situations.
  - Focus on production (throughput of submissions) at the expense of development of internal processes
  - High workload and subsequent deprioritization of improvement of internal processes
- Current controls:**
- A curation team member is in charge of documentation, both externally facing and internal.

**Estimated Risk:**

315 => 5 SEV \* 7 OCC \* 9 DET

## Individual Curator

*As part of the Review and decision on the removal request*

- Input:** Individual Curator
- Failure:**
- Selected the incorrect action to take based on the public status of the sequence records (intended to use withdraw rather than suppress)
  - Believed that the stored procedure command the curator intended to run (withdraw) was not working so tried to find a workaround in order to meet the request of the submitter
  - Relied on unofficial advice from other team members to determine an alternative course of action (using the kill command)
  - Used the wrong command to remove the sequence records given the existing 'public' status of the sequence records
- Impact:**
- A decision was reached that was not consistent with official INSDC policies that lead to the Curator using an inappropriate approach to removing the sequence records from public view
- Possible Causes:**
- Desire to be responsive to the request and needs of the submitter
  - High workload exacerbated by increasing submissions due to the pandemic



- Lack of knowledge of appropriate procedures for handling incoming requests (suppress vs kill)
  - Lack of comprehensive policies and procedures to guide the curator's behavior in this situation
  - Lack of clear escalation procedures that define situations that require an authoritative answer and how to obtain such an answer
  - Poor or missing feedback from the curation software to communicate the success/failure of attempts to run stored procedures
  - Curation interfaces (MS SQL Server Studio) that provide little to no user interface feedback to curators to guide their actions
- Current controls:**
- Some existing SOPs but this was not a topic covered directly.
  - Initial curator training (which may have been a number of years in the past)
  - Weekly curation team meetings to share knowledge
  - Ad hoc internal curation team discussions
  - The (b) (4) stored procedure does correct the effects of running withdraw on public data by enacting the suppress functionality instead; however, it does not appear to provide feedback to the curator that this has happened, nor does it have any similar checks and balances for Kill

**Estimated Risk:** 135 => 5 SEV \* 3 OCC \* 9 DET

#### Curation team knowledge

*As part of the Review and decision on the removal request*

- |                         |  |
|-------------------------|--|
| <b>Input:</b>           | <b>Curation team knowledge</b>   |
| <b>Failure:</b>         | <ul style="list-style-type: none"> <li>• The primary curator received incorrect advice from other team member(s)</li> </ul>  |
| <b>Impact:</b>          | <ul style="list-style-type: none"> <li>• A decision was reached that was not consistent with official INSDC policies</li> </ul>  |
| <b>Possible Causes:</b> | <ul style="list-style-type: none"> <li>• Lack of knowledge of appropriate procedures for handling incoming requests (suppress vs kill)</li> <li>• Lack of comprehensive policies and procedures to guide the curator's behavior in this situation</li> <li>• Lack of clear escalation procedures that define situations that require an authoritative answer and how to obtain such an answer</li> </ul> |

- Current controls:**
- Some existing SOPs
  - Weekly curation team meetings to share knowledge
  - Ad hoc internal curation team discussions

**Estimated Risk:** 96 => 4 SEV \* 4 OCC \* 6 DET

## 6.8. Step 8 - Action is taken based on the decision

### Individual curator

*As part of the Action is taken based on the decision*

**Input:** Individual curator

- Failure:**
- Curator attempted to run the (b) (4) command and determined it did not work
  - Curator then learned about the (b) (4) option as an alternative way to achieve the desired end result of removing the sequence records from SRA

- Impact:**
- A decision was reached to kill the sequences that was not consistent with official INSDC policies

- Possible Causes:**
- Lack of clear understanding of the INSDC status terminology (public, confidential, suppressed, replaced and killed)
  - No clear reporting process when the (b) (4) did not work
  - No clear documentation to guide the curator when the (b) (4) procedures did not work
  - Lack of training on the curation tools
  - Perceived urgency from the submitter
  - Lack of information on how to report issues with the (b) (4) command
  - High workload

- Current controls:**
- Initial curator training
  - Weekly curation team meetings

**Estimated Risk:** 324 => 9 SEV \* 4 OCC \* 9 DET

(b) (4) procedure

*As part of the Action is taken based on the decision*

<b>Input:</b>	(b) (4) procedure
<b>Failure:</b>	<ul style="list-style-type: none"><li>• - (b) (4) did not work, but (b) (4) did work</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• The stored procedure was not able to successfully run the withdraw command at the time of the incident so the kill command was run instead</li></ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• It is unknown at this time why the (b) (4) was unable to run with the (b) (4) parameter but was able to run with the (b) (4) parameter</li><li>• Curator unclear on INSDC terminology - did withdraw mean suppress or kill?</li><li>• MS SQLserver Studio provides little to no feedback to the curator about the results of running a stored procedure</li><li>• Curator has permissions to use the (b) (4) stored procedure with the kill parameter</li></ul>
<b>Current controls:</b>	<ul style="list-style-type: none"><li>• (b) (4) will automatically implement suppress if a curator tries to run withdraw on public sequences, however, there is no similar check if the curator runs kill on public sequences</li></ul>
<b>Estimated Risk:</b>	250 => 5 SEV * 5 OCC * 10 DET

## Curation interfaces

*As part of the Action is taken based on the decision*

<b>Input:</b>	<b>Curation interfaces</b>
<b>Failure:</b>	<ul style="list-style-type: none"><li>• The curator utilized the two curation interfaces to check to see if the (b) (4) command worked successfully (it did not)</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• No error messaging on the curator's user interface to let the curator know that public sequences can't be killed</li></ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• Two curation interfaces, MS SQL Server Studio, and several communication channels can lead to confusion and inefficiency</li></ul>
<b>Current</b>	<ul style="list-style-type: none"><li>• No relevant controls were mentioned in the interviews</li></ul>

controls:

**Estimated Risk:** 150 => 3 SEV \* 3 OCC \* 10 DET

## 6.9. Step 9 - Synchronization of SRA systems based on status change

### List of file locations for affected sequence records

*As part of the Synchronization of SRA systems*

<b>Input:</b>	<b>List of file locations for affected sequence records</b>
<b>Failure:</b>	<ul style="list-style-type: none"><li>• The files manually copied into the COVID bucket were not originally added to the list of file locations for the affected sequences</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• No action could easily be taken on these sequences as their locations were not tracked in the system</li></ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"><li>• Short timeframe to move COVID data to a new bucket in the cloud precluded the use or adaptation of existing processes</li></ul>
<b>Current controls:</b>	<ul style="list-style-type: none"><li>• No relevant controls were in place at the time the sequences were originally removed on 6/17/20</li></ul>
<b>Estimated Risk:</b>	<b>140</b> => 7 SEV * 2 OCC * 10 DET

### Policies and procedures for the cloud

*As part of the Synchronization of SRA systems*

<b>Input:</b>	<b>Policies and procedures for the cloud that define what needs to happen following a change in status</b>
<b>Failure:</b>	<ul style="list-style-type: none"><li>• Missing policies for the appropriate handling of sequence records in the cloud following withdrawal, suppress, or kill</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• With no definition of what should have happened to the files in the cloud it was not possible for any downstream processes to know what should happen in these cases and hence no action was taken</li></ul>

	on the files in any cloud bucket
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>No INSDC policies or best practices yet exist that reflect the idiosyncrasies of data handling in a cloud bucket</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>None</li> </ul>
<b>Estimated Risk:</b>	400 => 4 SEV * 10 OCC * 10 DET

## 6.10. Step 10 - Monitoring of overall system consistency

### Curation interface

*As part of the Monitoring of overall system consistency*

<b>Input:</b>	<b>Curation interface</b>
<b>Failure:</b>	<ul style="list-style-type: none"> <li>The curation interface only provides access to the curation history of a specific sequence, or set of sequences from a single submission</li> </ul>
<b>Impact:</b>	<ul style="list-style-type: none"> <li>It is a very manual process for SRA staff to track down or monitor the current status of a specific sequence or set of sequences, let alone the contents of the SRA database as a whole</li> </ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>The curation interface is focused on supporting curation tasks, often at a single sequence record level</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>Manual review of specific records</li> </ul>
<b>Estimated Risk:</b>	108 => 4 SEV * 3 OCC * 9 DET

### Sequence record status data

*As part of the Monitoring of overall system consistency*

<b>Input:</b>	<b>Sequence record status data</b>
<b>Failure:</b>	<ul style="list-style-type: none"> <li>Information about a sequence and its associated files are stored in multiple systems and locations</li> </ul>

<b>Impact:</b>	<ul style="list-style-type: none"> <li>• It is a very manual process for SRA staff to track down or monitor the current status of a specific sequence or set of sequences, let alone the contents of the SRA database as a whole</li> <li>• It is virtually impossible to monitor the entire SRA collection to detect situations where sequences are not replicated appropriately, or where backup copies may be missing</li> </ul>
<b>Possible Causes:</b>	<ul style="list-style-type: none"> <li>• Creation of an integrated view of the SRA sequence collection would be a significant software engineering undertaking</li> <li>• Generating and updating such a view would be challenging given the amount of data contained within SRA and its distributed nature</li> </ul>
<b>Current controls:</b>	<ul style="list-style-type: none"> <li>• Ad hoc execution of specialized database queries</li> </ul>
<b>Estimated Risk:</b>	<b>108</b> => 4 SEV * 3 OCC * 9 DET

## 6.11. Step 11 - Internal investigation into management of the sequences

In addition to the formal RCA, BioTeam was asked to consider NCBI/SRA's general response to the situation in the immediate period after they were notified of the potential issues regarding these sequences. The actions in this step (and in Step 12) did not directly contribute to the problem under consideration and so fall outside of the FMEA approach utilized in the previous sections, however, we present the findings for these sections using a similar format, focusing on the inputs and potential issues that we identified that made it harder for NCBI/SRA to respond to this situation.

### SRA Incident Response

#### process

*As part of the internal investigation*

<b>Input:</b>	<b>SRA incident response process</b>
<b>Issue:</b>	<ul style="list-style-type: none"> <li>• Existing procedures did not cover responses to incidents of this type</li> </ul>
<b>Impact:</b>	<ul style="list-style-type: none"> <li>• Ad hoc process had to be developed to respond to the incident and track down relevant information</li> <li>• Developing such processes and responding on the fly is not ideal as it is less efficient than following an established procedure, plus there is a higher risk of incorrect action or missed steps</li> </ul>

## Email communications

*As part of the internal investigation*

- |                |   |
|----------------|---|
| <b>Input:</b>  | <b>Email communications</b>   |
| <b>Issue:</b>  | <ul style="list-style-type: none"><li>• Potentially relevant Email conversations were spread across a variety of systems (MS Dynamics and personal email boxes)</li></ul> |
| <b>Impact:</b> | <ul style="list-style-type: none"><li>• Harder for SRA to track all communications on this topic to present a coordinated response</li></ul>                              |

## Applicable SRA policies

*As part of the internal investigation*

- |                |   |
|----------------|---|
| <b>Input:</b>  | <b>Applicable SRA policies</b>  |
| <b>Issue:</b>  | <ul style="list-style-type: none"><li>• Not documented in a single, authoritative location</li></ul>  |
| <b>Impact:</b> | <ul style="list-style-type: none"><li>• Applicable policies had to be manually brought together to present a coordinated response</li></ul> |

## Internal training

*As part of the internal investigation*

- |                |  |
|----------------|--|
| <b>Input:</b>  | <b>Internal training</b>   |
| <b>Issue:</b>  | <ul style="list-style-type: none"><li>• Not tracked or documented</li></ul>  |
| <b>Impact:</b> | <ul style="list-style-type: none"><li>• Unable to demonstrate previous training activities and curricula that could speak to the expected actions of the staff involved.</li></ul> |

## Internal policies

*As part of the internal investigation*

- |                |   |
|----------------|---|
| <b>Input:</b>  | <b>Internal policies</b>  |
| <b>Issue:</b>  | <ul style="list-style-type: none"><li>• Not documented in a single, authoritative location</li></ul>  |
| <b>Impact:</b> | <ul style="list-style-type: none"><li>• Lack of clarity about which version of which document was the most recent and/or was currently in effect.</li></ul> |

### Relevant data from the SRA database

*As part of the internal investigation*

<b>Input:</b>	<b>Relevant data from the SRA database</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Only partially accessible via existing web user interfaces</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Significant manual effort required to find and integrate available information to develop a picture of events</li></ul>

### Internal 'hold' on modifying related records within SRA

*As part of the internal investigation*

<b>Input:</b>	<b>Internal 'hold' on modifying related records within SRA</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Prevented updates or other modifications of related data until administrative reviews were completed</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Preserved the state of the system pending an external review</li><li>• Delayed any potential corrective actions</li></ul>

### Log files from relevant scripts and processes

*As part of the internal investigation*

<b>Input:</b>	<b>Log files from relevant scripts and processes</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Information distributed across multiple locations and sources</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Time consuming to acquire, integrate, and analyze raw data in order to gain an understanding of what happened</li></ul>

## 6.12. Step 12 - Public response by NIH based on data gathered by internal investigation

The actions in this step (and in Step 11) did not directly contribute to the problem under consideration and so fall outside of the FMEA approach utilized in the previous sections, however, we present the findings for these sections using a similar format, focusing on the



inputs and potential issues that we identified that made it harder for NCBI/SRA to respond to this situation.

## Official policies and procedures

*As part of the initial public response*

<b>Input:</b>	<b>Official policies and procedures</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Not easy to package and present to the outside world in a cohesive, coherent form</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Difficult for both scientists and the lay audience to understand what SRA does, what its governing policies are and what these mean in practice</li></ul>

## Information gathered about the incident

*As part of the initial public response*

<b>Input:</b>	<b>Information gathered about the incident</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Not easy to package and present to the outside world in a cohesive, coherent form</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• Disruption to the ongoing work at SRA as information is collected</li><li>• Difficult for external audiences (scientists and the lay audience) to clearly understand what is going on</li></ul>

## Communication Channels

*As part of the initial public response*

<b>Input:</b>	<b>Available communication channels - email, press releases, NCBI website(s), social media accounts</b>
<b>Issue:</b>	<ul style="list-style-type: none"><li>• Potential communication channels were not used</li></ul>
<b>Impact:</b>	<ul style="list-style-type: none"><li>• No additional information about the situation was available via relevant NCBI websites, or via web pages associated with relevant database records</li><li>• Visitors to these websites and/or searching for these specific records</li></ul>

were unable to determine what, if any, actions were taking place regarding these records

## 7. Root Causes

As stated in Section 2, **the core problem is that the sequences submitted by the group from Wuhan, China were not managed in a manner consistent with the current policies applicable to the NIH Sequence Read Archive.**

Inconsistencies with the current policies were found in three separate situations:

- 1. The public Wuhan sequences were killed rather than being suppressed**
  - The sequence records were public and should have been Suppressed in order to take them down from the public SRA site. However, the curator chose withdraw, which should not be used on already public records.
  - The curator then ran kill rather than withdraw because the withdraw command was found not to work.
- 2. Access to sequence files in the cloud is inconsistent with the current INSDC policies**
  - Sequences are still accessible via accession number even though they were killed which should result in them being inaccessible via accession. (IT systems)
  - There is no accepted policy or procedure for how to take files offline from the cloud. (Policies and procedures)
- 3. 18 source data files have been lost which does not meet NLM's goal of preserving its collected data**
  - Storage shortage. (IT Systems, Organization/budget)
  - Errors during rapid migration to the cloud. (IT Systems, Organization/budget)

Based on our analysis of the information provided to BioTeam through interviews and ancillary documentation, we have identified a variety of root causes that contributed to the problem. We describe these for the three areas of inconsistency described above and then provide an overall view of how these and other factors ultimately contributed to the situation at hand.

### 7.1. Situation 1 - The public Wuhan sequences were killed rather than being suppressed

There were two contributory elements to this situation:

1. The sequence records were public and should have been Suppressed to take them down from the public SRA site. However, the curator chose to use 'withdraw' which should not be used on already public records.
2. The curator then ran kill rather than withdraw because the withdraw command was found not to work.

The direct root causes are as follows:

- Communication within the SRA Operations Team
  - The curator did not have a solid understanding of how to proceed in this situation: which action to take to manage the remove request appropriately, how to respond when the withdraw command failed, what additional actions should be taken given the fact that these were sequences of high interest (COVID-19 sequences potentially relevant to the global pandemic).
- Inadequate Policies and procedures
  - The existing policies and procedures did not adequately discuss status change options, nor did they adequately relate SRA status change options to overarching INSDC policies.
  - They did not adequately define an official escalation or review process that should be applied to validate potentially high impact actions.
  - SRA's policies do not have provisions for extraordinary situations (such as a global pandemic) that may require specific modifications to standard policies and procedures.
- IT Systems
  - The specific command that the curator wanted to run, failed with little or no feedback to the curator.
  - There were no system-level checks and balances to prevent the curator from running a command that was inappropriate to use on public sequence data.

A variety of systemic factors contributed to these root causes:

- Reduced focus on developing and maintaining policies and procedures, due to prioritization of more urgent, day-to-day production-level activities.
- Reduced focus on training, due to assumptions about curator knowledge based on past performance and length of time in the role, and prioritization of more urgent, day-to-day production-level activities.
- Reduced focus on improvements to internal curation systems and tools, due to prioritization of more urgent, day-to-day production-level activities.

## 7.2. Situation 2 - Access to sequence files in the cloud is inconsistent with the current INSDC policies

There were two contributory elements to this situation:

1. Sequences are still accessible via accession number even though they were killed which should result in them being inaccessible via accession.
2. There is no accepted policy or procedure for how to take files offline from the cloud.

The direct root causes are as follows:

- IT Systems
  - The processes that would normally keep the SRA storage systems in sync with the current status of individual sequence records did not work with the sequences stored in any cloud bucket at the time of this incident. Permissions for any on-prem sequences can be adjusted to make specific sequence files inaccessible to be in line with that sequence's INSDC status, however, a comparable process had not been implemented for the sequences stored in the cloud.
  - The reason for this is that the architecture of the cloud is sufficiently different from on-premises systems that the existing approaches do not apply and there were no policies or procedures yet available to provide guidance on how to manage sequence data in the cloud in a manner which is compatible with current INSDC policies.
- Inadequate Policies and Procedures
  - SRA is one of the first INSDC databases to move significant portions of their repository into the cloud. As such, SRA is one of the first to experience the challenges associated with moving such large datasets (40PB) into a cloud environment and managing the individual files and associated metadata using the cloud-native tools provided by each cloud service provider.

A variety of systemic factors contributed to these root causes:

- Greatly increased focus on moving data and systems to the cloud at SRA, due to desire to reduce costs and take advantage of the cloud and the NIH STRIDES program.
- Challenges adapting traditional on-premises data handling approaches to a cloud environment due to the differences and added complexities present in the cloud as a whole, and as a result of supporting multiple cloud providers.

### 7.3. Situation 3 - 18 source data files have been lost, which does not meet NLM's goal of preserving its collected data

There were two contributory elements to this situation:

1. A shortage of on-premises storage.

2. Errors that occurred during the rapid migration of on-premises data into the cloud.

The direct root causes are as follows:

- IT Systems
  - SRA ran out of on-premises storage space in April 2020.
  - The scripts developed to move large volumes of data from the on-premises storage into the cloud were implemented very rapidly and utilized specific cloud components (spot instances) to manage the costs associated with such a huge data migration. However, due to the complexity of the task being undertaken there were some uncaught errors that resulted in certain files not being copied into the cloud. When the on-premises backups of the source data files were deleted to free up space, the only remaining copies of these files that had not been successfully copied to the cloud were lost.

A variety of systemic factors contributed to these root causes:

- Greatly increased focus on moving data and systems to the cloud at SRA, due to desire to reduce costs and take advantage of the cloud and the NIH STRIDES program.
- Limited storage capacity on-premises due to prioritization of moving to the cloud over continued expansion of on-premises infrastructure.
- The pandemic negatively impacted procurement and shipping of new storage hardware.
- Challenges adapting standard data handling approaches to work with a cloud environment due to complexities associated with moving such large volumes of data into a cloud environment in a very short timeframe.

## 7.4. Systemic Root Causes

The sections above (7.1 – 7.3) describe specific examples of the overall problem statement, indicating where the sequences in question were not managed in accordance with the current policies applicable to SRA, and describe the direct root causes that contributed to each example. Our analysis also identified several underlying, systemic root causes that set the stage for the specific issues identified above (these are listed in Table 3 below).

Systemic Root Cause	Description	Also impacted by
COVID-19 pandemic	The pandemic affected the situation in several significant ways. Like many other groups at NIH and around the world, SRA was working hard to respond to the pandemic and support the scientific community, which added to the general sense of urgency at SRA. The role of SRA in such a public health emergency is not clearly defined by NIH and is a significant departure	

	from the normal, more academic and research-centric sequence submissions and inquiries SRA normally handles. Clearly the nature of the sequences in question and their potential value in the study of SARS-CoV-2 was a factor in the exposure generated by this situation and was not a situation for which SRA had existing policies and procedures in place to handle. The pandemic additionally lengthened procurement times for new infrastructure, impacting SRA's ability to acquire new hardware as storage space ran low.	
<b>Budget</b>	Submissions to SRA continue to increase and expectations of SRA continue to increase, however the budget has not kept pace, impacting such things as SRA's ability to: hire additional staff, maintain adequate infrastructure, develop and maintain internal systems, and create and update policies and procedures.	
<b>Reactive focus vs. proactive</b>	SRA's focus is on urgent & important tasks (submissions, value-add features, lack of storage, cloud migration), to the detriment of non-urgent but important tasks (policies, procedures, training, curation interfaces, documentation).	Budget, Staffing, Workload, SRA priorities
<b>SRA strategic goals</b>	The current strategic goals (Strategy Documents) focus on 1) Sustainability, 2) Enabling users to find the data they need, and 3) Integrating SRA with the emerging NIH data ecosystem. These are very appropriate functional goals for a platform such as SRA, however, the focus is on externally facing goals - the data, infrastructure, and features. There are no internal goals to address training, communication, development of procedures and policies which greatly contribute to the ability of SRA to provide a robust platform to serve the scientific community.	Budget, Staffing, Workload
<b>Lack of diversity</b>	SRA serves a diverse global community. Therefore policies, procedures, and ultimately staff should reflect this global diversity. Out of the 18 people we interviewed there were 0 people of color and only 4 women (2 that interface with the submitters). This lack of diversity is most likely negatively impacting internal and external communication, innovation, as well as the ability of the SRA staff to create policies and procedures for a global user base.	SRA priorities

*Table 3 - List of systemic root causes*

## 8. Future Considerations

The root cause analysis presented in the prior sections clearly indicates that specific breakdowns in various aspects of SRA operations and functions, critical infrastructure inadequacies, budgets that don't match production expectations, pressure to migrate to the cloud in a haphazard manner, as well as lack of clarity and guidance for SRA on the topic of public health emergencies, led to the withdrawal of the Wuhan sequences in question as requested by the submitting researcher yet left the data findable in the SRA cloud archives.

**Importantly, we find no obvious mal intent behind the actions taken during this incident on the part of SRA or its staff. We find a series of weaknesses in the system that led to this situation during an unprecedented public health emergency that has also been colored by political polarization on the topic of the COVID-19 pandemic origins.**

In this section, we indicate potential opportunities for mitigation of the failure modes identified in the RCA Failure Modes and Effects Analysis (FMEA) outlined in Section 6 above. These opportunities are meant to act as a guide and to show the probable impact the prior mitigation of each issue might have had on the outcome of this incident.

The FMEA analysis includes an analysis of "Potential Opportunities" (see RCA table in the attached supporting document, "Potential Opportunities" column). Per that analysis, below are process inputs that are assumed / identified as contributors to the problem. Additional details are in the RCA table, but an exploration of the potential opportunities for NLM's consideration are listed in this section.

### 8.1. Evaluating potential opportunities

The first step in analyzing potential opportunities to mitigate the issues identified in the previous sections of this document is to look at the highest value targets as ranked by the Risk Priority Number (RPN) in the RCA outlined in Section 6. In order to prioritize which targets to focus on for opportunities, each category identified in the RCA Fishbone analysis (Figure 4 4) was assigned as a contributing factor to process steps identified in Section 6 and each process input's RPN was summed across each of those categories.

The total RPN scores per category are shown in Figure 7 below.



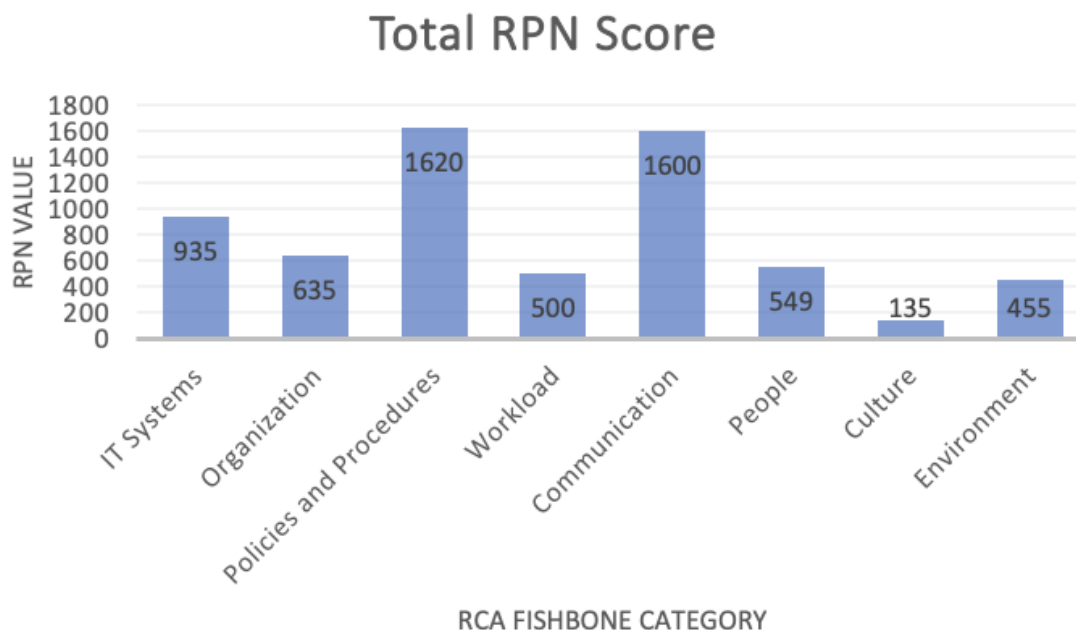


Figure 7 7 - Total RPN scores for the identified fishbone categories outlined in Section 6.

While the overall scores per category are useful on their own to distinguish the categories that most contributed to the current situation, those RPNs are made up of three components: severity, occurrence, and detection (see Section 6 for more details). As such, it is useful to see how the components contributed to the total score per category.

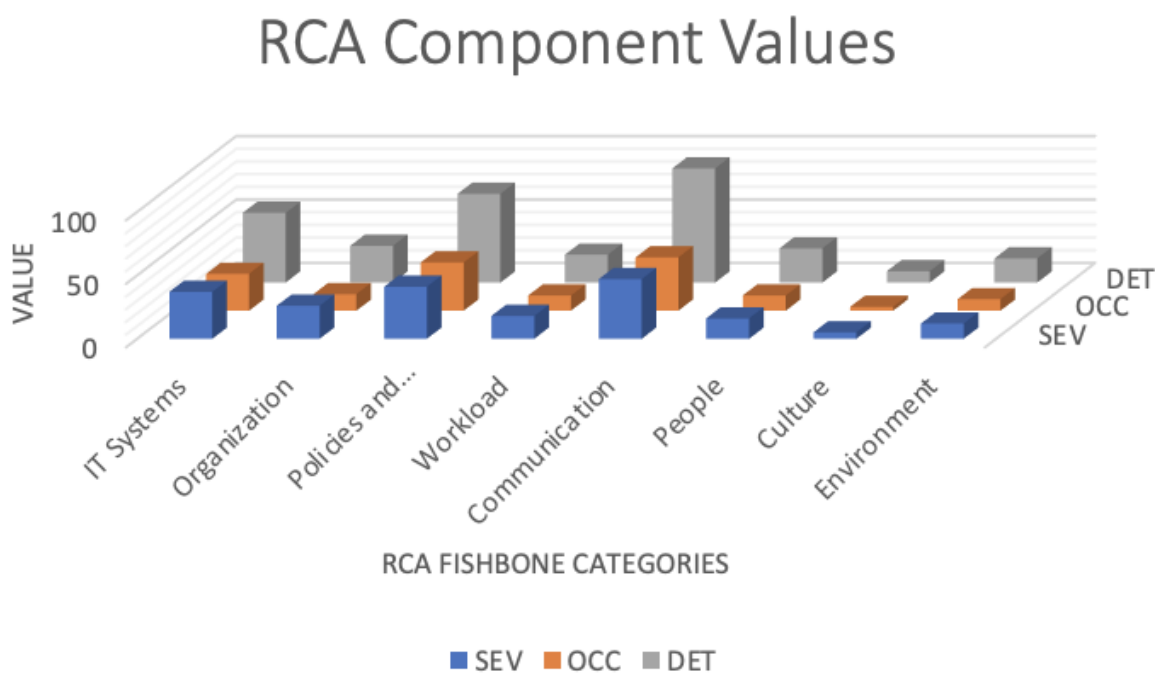


Figure 8 8 – RCA Component Values / RCA Fishbone Categories

While the overall scores per category are useful on their own to distinguish the categories that most contributed to the current situation, those RPNs are made up of three components: severity, occurrence, and detection (see Section 6 for more details). As such, it is useful to see how the components contributed to the total score per category.

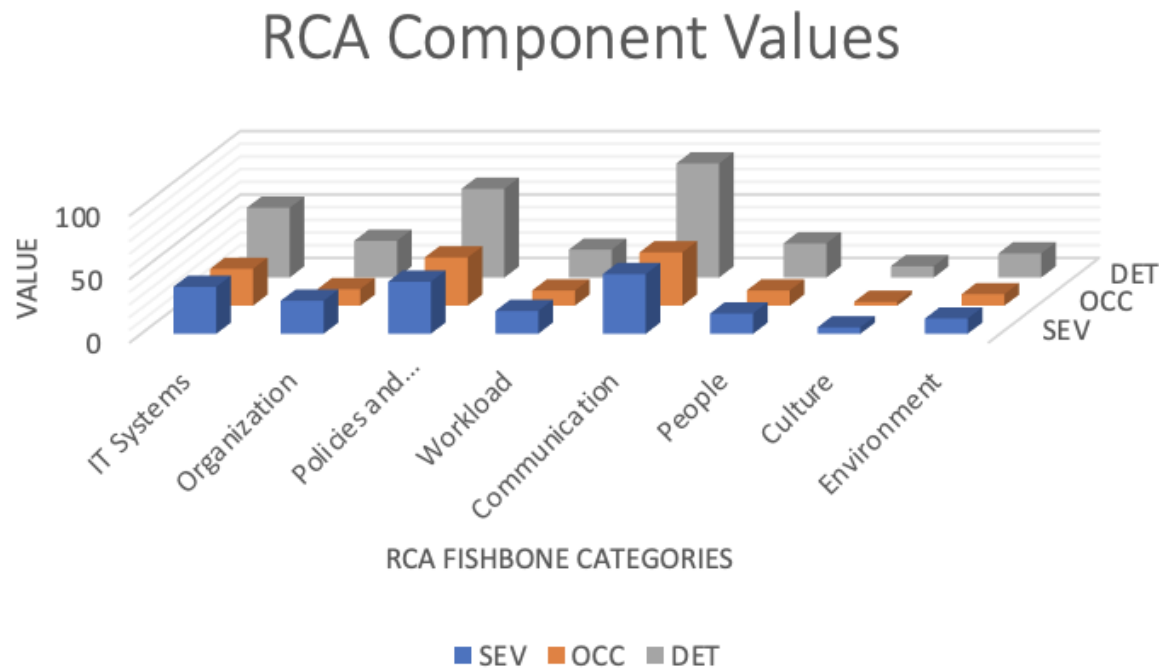


Figure 8 above shows a similar distribution as the total RPNs do in Figure 7. While the severity of failure and the frequency of occurrence scores are highest for Communication and Policies & Procedures, the risk of not detecting the problem for those categories and IT Systems are very high. This indicates that the events that led to the current situation were less likely to be detected by the SRA team than some of the other categories that either had lower impact or had more controls or awareness associated with them. It also indicates that these categories the highest priority to fix and that doing so would have a positive impact on SRA.

Like the RCA component value figure above, Figure 7 clearly shows two distinguishable modalities within the total RPN scores. If the categories are split into tier 1 and tier 2 opportunities along that modality, we can establish that tier 1 opportunities would represent the highest impact considerations going forward for SRA to focus on, with a lower potential impact (on varying levels) for tier 2 opportunities.

#### **Tier 1 Opportunities**

1. Policies & Procedures
2. Communication
3. IT Systems

#### **Tier 2 Opportunities**

4. Organization
5. People
6. Workload
7. Environment
8. Culture

## 8.2. Tier 1 Opportunities

The following opportunities address the actions around creating more consistent and forward-looking policies and procedures, changing the model of communication within the team, and updating the technological backlog and infrastructure of IT systems.

Each item below is marked with a general indicator as shown in Table 4, below.

Factor for consideration	Abbreviation	Indicator Level and Color		
Level of effort to implement	LOE	Low	Medium	High
Overall impact of implementation	IMP	High	Medium	Low
Risk if not implemented	RISK	Low	Medium	High

*Table 4 - Factors for consideration when reviewing identified opportunities*

The scale for each indicator is on a simple scale of Low, Medium, or High relating to the indicator against each opportunity. Note that the scale for impact is reversed from level of effort and risk, since high impact is a positive identifier.

### 8.2.1. Policies and Procedures

The following potential opportunities may address the largest issues that led to the Wuhan SARS-CoV-2 sequence issues that resulted in the current situation. In the list below, we indicate which policies and procedures could be examined to improve the overall quality of operations, service, and reliability of SRA and would likely have reduced the risk for this incident occurring in the first place.

#### Cloud Storage Policy:

LOE	IMP	RISK
-----	-----	------

- Create a cloud data synchronization policy that describes how SRA data stored in cloud locations will be updated when the status changes in the SRA database.
- Create a cloud data-retention policy that outlines the types of data, the availability zones, and the data-type retention times and locations (cloud or on-premises archive) of data backups.

- 
- Define an internal SRA policy that governs the structure of cloud storage buckets, their security settings, and the situations in which those settings are to be changed and by whom.

---

**INSDC Policy:**
**LOE**
**IMP**
**RISK**

- Update internal SOPs to better describe the role of the INSDC policies, and how they are implemented at SRA.
- Update curation software, command names, and parameters to use only the appropriate INSDC status terms.
- Institute shared, recurrent training with other NIH INSDC databases (e.g., GenBank) to ensure consistent interpretation and application of INSDC policy across all relevant NIH databases.

---

**Internal Procedures:**
**LOE**
**IMP**
**RISK**

- Create a policy that clearly outlines thresholds of decision points for modification of SRA records and data and chain of escalation and approval for those decisions.
- Update internal documentation to address the missing content.
- Institute regular review and update processes to keep documentation content current.
- Institute regular training to ensure curators are correctly following the existing policies and procedures and are also familiar with any changes and additions.
- Ensure sufficient time is available for staff members to work on documentation.
- Define and track internal curation-related goals and metrics (e.g., docs should be reviewed and updated every quarter or following any 'incident' (conditions TBD)).
- Develop definitions for what constitutes 'extraordinary situations' that SRA may be facing, develop and implement corresponding processes for acknowledging and responding to these situations.
- Implement a feedback and feature request system that is communicated across SRA teams.

---

**NIH-Level SRA Policies:**
**LOE**
**IMP**
**RISK**

- Request guidance from NIH Office of the Director (OD) that defines how SRA data is to be kept with regards to federal records management policies.
  - Request guidance from NIH OD that defines whether SRA should treat public health-related data vs. academic voluntary data differently.
  - Request guidance from NIH OD that defines the operational status, activities, thresholds, and modification to standard operations when SRA data is deemed critical to a public health emergency (e.g., COVID-19 global pandemic).
- 

## 8.2.2. Communication

The following potential opportunities may address the largest issues that led to the Wuhan SARS-CoV-2 sequence issues that resulted in the current situation. In the list below, we indicate which communications practices and improvements could be examined to improve the overall quality of operations, service, and reliability of SRA and would likely have reduced the risk for this incident occurring in the first place.

---

**The Submitter's Email and Subsequent Communications with the Submitter:****LOE****IMP****RISK**

- Clearly define the terms and conditions for submitting data to SRA so that the end user is aware of the circumstances in which data can be killed, withheld, or released.
- Update Standard Operating Procedures (SOPs) to provide clearer instructions on how and when to use templates for user communications.
- Institute regular recurrent training, and the use of templates and the goals behind their use.
- Institute periodic 'spot checks' or reviews of curator responses to submitters.

---

**Individual Curator:****LOE****IMP****RISK**

- Provide immediate refresher training on key curation topics such as INSDC and the appropriate methods of handling status changes.
- Institute regular refresher training and updates for all SRA staff on features and bugs in the SRA system.
- Institute a recurring training program to ensure key topics are reviewed by the SRA team (curators and developers).
- Improve the curation tools such that incorrect status changes are much harder to perform like:
  - Feedback to curator prior to running the command
  - 'Dry run' to show the effects before the command is run
  - Checks in the code itself to ensure business logic is followed
  - Requirement for two-step approval for high impact actions
  - Alerts to relevant higher-level SRA staff following the use of certain high-impact commands
- Provide training on using stored commands and implications for their use.

---

**Public relations:****LOE****IMP****RISK**

- Place the official NIH response to the Wuhan incident on the BioProject page (and associated pages) so when people look for the data/response it is clear that an investigation is ongoing.
  - Publish an updated official statement with the results of this investigation. Outline the steps that have been taken and will be taken to prevent future misunderstandings around status change requests of sequences.
  - Add the updated official statement to the comments section of Bloom preprint on bioRxiv.
  - Tweet a link to the updated official statement either from an SRA twitter account or the NIH twitter account.
  - Release the 241 sequences in question since they have been published.
  - Clearly define sequence status states on the SRA FAQ pages and what actions will be taken when submitters use certain words.
  - Clearly state no data is/was deleted.
  - Develop an SRA user agreement (to be signed upon submission) with clear definitions of terms for status changes and also make it very clear sequences are being copied to INSDC partners and also cloud environments and will not be permanently deleted unless the data contains identifiable human data.
  - Ask the community for help/feedback when developing the user documentation to ensure that inclusion of diverse user perspectives is represented in the SRA documentation.
-

---

**SRA Accession Numbers:****LOE****IMP****RISK**

- Provide detailed documentation to users and curators around these various assigned numbers (SRA accession numbers; BioProject numbers; and BioSample numbers.)
- 

### 8.2.3.IT Systems

The following potential opportunities may address the largest issues that led to the Wuhan SARS-CoV-2 sequence issues that resulted in the current situation. In the list below, we indicate which IT systems, SRA features, and activities could be examined to improve the overall quality of operations, service, and reliability of SRA and would likely have reduced the risk for this incident occurring in the first place.

---

**Microsoft Dynamics Improvements:****LOE****IMP****RISK**

- Improve the operations of MS Dynamics to effectively function as a well-aligned ticketing, history, and incident response system for SRA.
  - Improve MS Dynamics functionality to make use of templates easier.
  - Improve MS Dynamics to incorporate more formal review/approval workflows that ensure SOPs are followed.
- 

(b) (4)

**Procedure:****LOE****IMP****RISK**

- Integrate the change of status commands into the curator interface and remove access to stored procedures in the database.
  - Review process for who has permission to run the kill command.
  - Add logic into the procedure (or similar) to prevent public sequences from being 'killed' without any additional checks and balances.
  - Better error messaging when scripts don't work (i.e., tell the curator (b) (4) didn't work because it is a violation of the INSDC protocols to kill a public sequence).
- 

**Curation Interfaces:****LOE****IMP****RISK**

- Create one web-based curator interface.
  - Review of curator usage of xxx commands or build their function into the curator interface and remove direct access to the stored procedures in the database.
- 

## 8.3. Tier 2 Opportunities

The Tier 2 opportunities listed in this section didn't have as much of an impact on the events that led to the RCA of the Wuhan sequences. While they are lower impact, they are important factors to consider in future planning and improvements to the SRA system and operations.

Here we use the same rating system used in Section 8.2, but a lighter format for the presentation of the opportunities, since they aren't as detailed as the Tier 1 items.

### 8.3.1. Organization

LOE	IMP	RISK
-----	-----	------

Organization refers to the structures that support SRA, which include how teams are organized and managed, how checks and balances are established, the political climate, economic pressures, government incentives, new technology, leadership involvement, and group norms.

Identified potential opportunities:

1. Establish checks and balances or QA/QC steps for operations.
2. Develop oversight into SRA by NLM as a whole.
3. Create knowledge redundancy within the organization.
4. Work to reduce the siloes between the NCBI databases
5. Provide organizational structure to help staff know who knows what subset of information.

### 8.3.2. People

LOE	IMP	RISK
-----	-----	------

The people category refers to the human factors that influence the operations and functioning of SRA and that contributed to the problem. These factors include staffing, diversity, training, and expectations. This category also considers aspects related to individual knowledge and capabilities, group norms, knowledge and skills, experience, and motivation/attitude. Identified potential opportunities related to people are:

1. Create skill redundancy within the team.
2. Develop additional training for operations and developers to tackle new demands of the SRA platform.
3. Develop user facing documentation and user agreements with inputs from a diverse set of voices.
4. Increase diversity of SRA staff to help match the diversity of the user base.

### 8.3.3. Workload

LOE	IMP	RISK
-----	-----	------

This category includes two aspects of how the workload effects SRA's operations; 1) the amount of work relative to the number of staff and, 2) how that work is prioritized for the staff. Workload also considers the nature of the work, the complexity of the work, the focus on individual vs teamwork, and resource availability. Identified potential opportunities related to workload are:

1. Acquire a new ticketing system with input from the SRA staff
2. Ensure organizational priorities (such as moving to the cloud, making COVID-19 data more accessible, launching new features, and increasing customer ease of access to

data) align with the developer team's capacity and the operation team's ability to understand and support them.

3. Prioritize development time for the team to finish things that are undone and need to for long-term internal success.

#### 8.3.4.Environment

LOE	IMP	RISK
-----	-----	------

Environment refers to the totality of external factors that were happening during the time of the incident that contributed directly or indirectly to the problem. Identified potential opportunities related to the SRA environment are:

1. Create a sustainability plan from a budgetary and operational perspective that also aligns with opportunities pointed out in Sections 8.2 and 8.3.
2. Reduce ongoing pressure to move SRA data to the cloud due to budgetary constraints by defining a clear sustainable path for growth and management of SRA data.
3. As a part of the sustainability plan, refactor budgets and infrastructure plans to include maintenance and upgrades of on-premises systems as they are critical to the function of SRA.
4. Proactively work to prepare for the next pandemic (and associated data) to avoid the pitfalls associated with this case.

#### 8.3.5.Culture

LOE	IMP	RISK
-----	-----	------

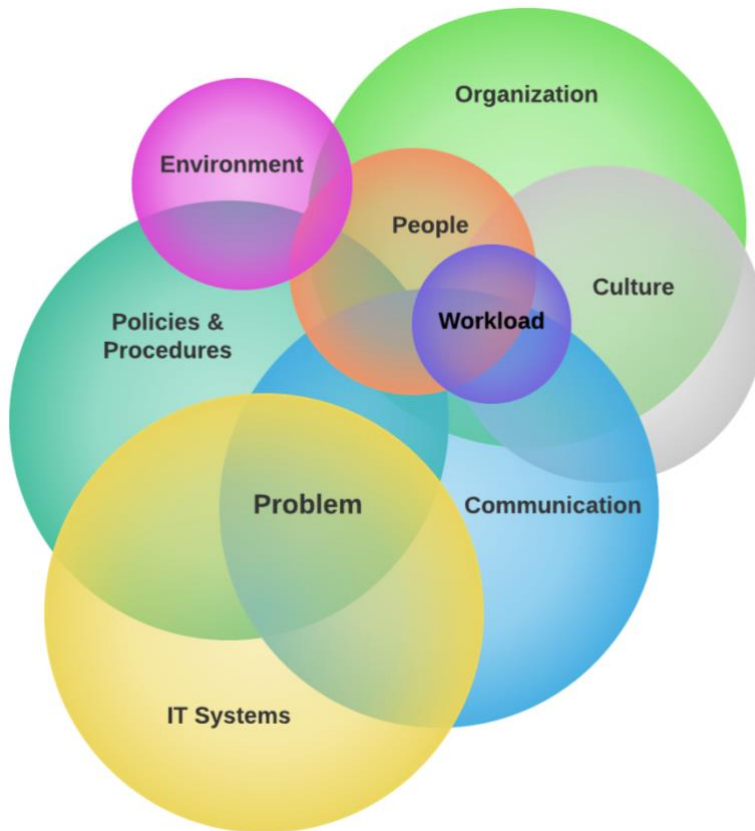
Culture refers to the diversity, demographics, cultural competency, internal understanding of incentives, transparency, voluntary error reporting, information sharing, and a willingness to follow an internal set of standards. Identified potential opportunities related to culture are:

1. Develop a culture of being proactive toward addressing the growing list of technical tasks (versus reactive) that need to be accomplished and updated.
2. Continuing to encourage the culture of caring that already exists within the organization.
3. Clarify the SRA mission to include the distinction between the initial motivation of SRA as an academic archive to one of supporting public health into the future.
4. Align the bigger picture metrics outlined in the strategic plan of the SRA (being in the cloud and having the data easily accessible) with the many other tasks that are not prioritized officially but rather prioritized internally (i.e., quick turnaround times for addressing tickets).



## 9. Conclusion

This report and root cause analysis were performed in response to the incident involving the removal of SARS-CoV-2 sequences from SRA at the request of the researcher from Wuhan, China in June 2020, but left the source sequences available in the cloud. This RCA was undertaken as a method of ascertaining the problem that led to this incident and to identify the causes associated with the problem. The goal of this analysis was to help the stakeholders understand the problem causes well enough to subsequently direct a plan of action towards the resolution of those problems.



The RCA undertaken in this report relates to the problem stated as:

“The sequences submitted by the group from Wuhan, China were not managed in a manner consistent with the current policies applicable to the NIH Sequence Read Archive.”

The causes were identified to be at the intersection of the lack of consistent internal and external policies and procedures, gaps in communication, and IT infrastructure deficiencies. Contributing, but not at the root of the problem were issues associated with the global pandemic, issues related to workload, culture, and organizational structure.

This RCA was concluded on August 12<sup>th</sup>, 2021.

# 10. Appendix

The following sections contain information relevant to this assessment.

## 10.1. NLM SRA RCA Interviewees (Anonymized)

19 total interviews were conducted from July 20th - July 29th, 2021 with a total of 18 interviewees (one interviewee was interviewed twice). For the purposes of this report, interviewee names will remain anonymous and will be referred to according to the following aliases throughout the report if referenced at all:

Leadership / Team Lead	Operations	Developer
L1	O1	D1
L2	O2	D2
L3	O3	D3
L4	O4	D4
L5	O5	
L6	O6	
L7	O7	

## 10.2. Scoring tables used in the FMEA

### 10.2.1. Severity

For issues related to data, rather than infrastructure (Code, hardware), severity is related to public expectations, scientific norms, INSDC policies, and then looks at the 'blast radius' for issues and finally for the more serious issues reflects problems related to data protection and privacy as the top end of the data-related severity scale.

Ranking	Severity Criteria
1	No effect
2	Very minor deviation from expected norms
3	Minor deviation from expected norms
4	Small deviation from expected norms The data/system does not require remediation.
5	Moderate deviation from expected norms The primary data is correct but some other aspect does require remediation (e.g. metadata)

6	Data handling differs significantly from expected, impacts mostly limited to internal/NIH stakeholders, e.g. data is still correct but is handled in a way that is inconsistent with internal policies
7	Data handling differs significantly from expected, impacts extend to immediate external stakeholders (submitters, other research users of the system), e.g. data is still correct but is handled in a way that is inconsistent with INSDC (or other relevant) policies
8	Data handling differs significantly from expected, impacts extend to a broader spectrum of external stakeholders (Researchers using the data, NIH leadership, the scientific community as a whole, the general public) e.g. data is incorrect, or missing, and has/could impact downstream analyses
9	Failure involves inadvertent release of data (e.g. Human data, or unpublished data) that might violate legal terms, and/or be in non-compliance with relevant NIH/Federal regulations or standards. Affected data is very limited and remediation can be completed very quickly
10	Failure involves inadvertent large scale release of data (e.g. Human data, or unpublished data) that violates legal terms, and/or puts the system into non-compliance with relevant NIH/Federal regulations or standards, and which can not be easily rapidly remediated, potentially suspending operation of the resource.

### 10.2.2. Occurrence

Ranking	Estimated frequency of occurrence
1	every few years or so...
2	Yearly
3	every few months
4	every month
5	every few weeks (2-3/mo)
6	every week
7	every few days (2-3 times a week)
8	every day
9	Multiple times a day
10	This is currently a permanent state of affairs

### 10.2.3. Detection

Ranking	Estimated probability of detection
1	Almost Certain Detection
2	Very High Chance of Detection
3	High Probability of Detection
4	Moderately High Chance of Detection

5	Moderate Chance of Detection
6	Low Probability of Detection
7	Very Low Probability of Detection
8	Remote Chance of Detection
9	Very Remote Chance of Detection
10	Absolute Uncertainty – No Control

### 10.3. Documents List / Reference Materials

#### Document Categories for materials provided by NIH for this RCA

Document Category	Short Name	Relevance to RCA
Analytics run on the Wuhan sequences by SRA	Analytics Documents	Moderate
Log files and SRA tracking records	Log files	High
Internal and external communications and presentations	Comms Documents	High
Policies and Procedures, training materials	Policy Documents	High
Strategy, status, alignment, and internal presentations and communications materials	Strategy Documents	Moderate
SRA user documentation and training materials	Documentation	Moderate

Alternate Name	Documentation Name	Date Reviewed
<b>Public Documents / Web References</b>		
Farkas 2020	Farkas, C., F. Fuentes-Villalobos, J. L. Garrido, J. Haigh, and M. I. Barría, 2020 Insights on early mutational events in SARS- CoV-2 virus reveal founder effects across geographical regions. <a href="#">PeerJ 8: e9255</a>	Downloaded July 23 <sup>rd</sup> , 2021
Wang 2020b	Wang, M., A. Fu, B. Hu, Y. Tong, R. Liu, et al., 2020b Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. <a href="#">Small 16: 2002169</a> .	Downloaded July 23 <sup>rd</sup> , 2021

Wang 2020a	Wang, M., A. Fu, B. Hu, Y. Tong, R. Liu, et al., 2020a Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. <a href="https://doi.org/10.1101/2020.03.04.20029538">medRxiv 10.1101/2020.03.04.20029538</a> .	Downloaded July 23 <sup>rd</sup> , 2021
Wang 2020b Sup Data	<a href="#">Supplemental Data from Wang 2020b</a>	Downloaded July 23 <sup>rd</sup> , 2021
Farkas 2020 Sup Table	<a href="#">Supplementary Table from the Farkas 2020</a> <a href="https://web.archive.org/web/20210502130356/https://dfzljdn9uc3pi.cloudfront.net/2020/9255/1/Supplementary_Table_1.xlsx">https://web.archive.org/web/20210502130356/https://dfzljdn9uc3pi.cloudfront.net/2020/9255/1/Supplementary_Table_1.xlsx</a>	Downloaded July 23 <sup>rd</sup> , 2021
Bloom 2021a	Bloom, J. 2021a. Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. bioRxiv <a href="https://doi.org/10.1101/2021.06.18.449051">https://doi.org/10.1101/2021.06.18.449051</a> version 1 posted June 22, 2021.	Added to our repository July 30 <sup>th</sup> , 2021
Bloom 2021b	Bloom, J. 2021b. Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. bioRxiv <a href="https://doi.org/10.1101/2021.06.18.449051">https://doi.org/10.1101/2021.06.18.449051</a> version 2 posted June 29, 2021	Added to our repository July 30 <sup>th</sup> , 2021
Bloom 2021c	Bloom, J., Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic, <i>Molecular Biology and Evolution</i> , 2021, <a href="https://doi.org/10.1093/molbev/msab246">https://doi.org/10.1093/molbev/msab246</a>	Last viewed, August 20 <sup>th</sup> , 2021
INSDCPol	INSDC Policy Statement: <a href="https://www.insdc.org/policy.html">https://www.insdc.org/policy.html</a>	Last referenced August 4 <sup>th</sup> , 2021
<b>BioTeam Created Documents Submitted with the Final Report</b>		
RCA Table	NLM SRA Root Cause Analysis Table - Final	Created; Ongoing Throughout the Project