NIST Cybersecurity White Paper
NIST CSWP 35 ipd

# Cybersecurity Threat Modeling the Genomic Data Sequencing Workflow

*An example threat model implementation for genomic data sequencing and analysis*

Ronald Pulivarti
*National Cybersecurity Center of Excellence*
*National Institute of Standards and Technology*

Justin Wagner
Justin Zook
*Material Measurement Laboratory*
*National Institute of Standards and Technology*

Brett Kreider
Julie Snyder
Kevin E. Wilson
Martin Wojtyniak
*The MITRE Corporation*

Scott Ross
Philip Whitlow
*HudsonAlpha Institute for Biotechnology*

Einaam Alim
Isabelle Brown
Patrick Pape
Jared Sheldon
*The University of Alabama in Huntsville*

This publication is available free of charge from:
https://doi.org/10.6028/NIST.CSWP.35.ipd

December 16, 2024

**NIST Technical Series Policies**



**How to Cite this NIST Technical Series Publication:**

Pulivarti R, Wagner J, Zook, J, Kreider B, Wilson K, Snyder J, Wojtyniak M, Ross S, Whitlow P, Sheldon J, Brown I, Pape P, Alim E (2024) Cybersecurity Threat Modeling the Genomic Data Sequencing Workflow: An example threat model implementation for genomic data sequencing and analysis. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 35 ipd.
https://doi.org/10.6028/NIST.CSWP.35.ipd

**Author ORCID iDs**

Ronald Pulivarti: 0000-0002-8330-3474
Justin Wagner: 0009-0003-8903-0504
Justin Zook: 0000-0003-2309-8402
Brett Kreider: 0009-0004-1508-5876
Julie Snyder: 0009-0004-6352-2831
Kevin Wilson: 0009-0008-3673-6040
Martin Wojtyniak: 0009-0005-9643-2194
Mohammad Einaam Alim: 0009-0005-3370-907X
Isabelle Brown-Cantrell: 0009-0004-8820-6448
Dr. Patrick Pape: 0009-0005-4922-4026
Scott Ross: 0009-0002-8672-6496
Jared Sheldon: 0009-0009-7909-4217
Philip Whitlow: 0009-0000-7677-3825

**Public Comment Period**

December 16, 2024 – January 30, 2024

All comments are subject to release under the Freedom of Information Act.

**Submit Comments**

Please submit comments to genomic_cybersecurity_nccoe@nist.gov with "Comments on NIST CSWP 35" in the subject field. We encourage you to use this comment template.

National Institute of Standards and Technology
Attn: Applied Cybersecurity Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 2000) Gaithersburg, MD 20899-2000

**Additional Information**

Additional information about this publication is available at <CSRC link>, including related content, potential updates, and document history.

**All comments are subject to release under the Freedom of Information Act (FOIA).**

## Abstract

Advancements in genomic sequencing technologies are accelerating the speed and volume of data collection, sequencing, and analysis. However, this progress also heightens cybersecurity and privacy risks. In this paper, the National Cybersecurity Center of Excellence (NCCoE) Genomic Data project team demonstrates a cybersecurity threat modeling using an example workflow involving an organization sending a physical sample to a genomic sequencing provider, then receiving back and processing the genomic data. This paper provides an example of how to conduct cybersecurity threat modeling, including documenting the architecture, identifying threats, applying sample mitigations, and iterating the process as needed. While this paper focuses on cybersecurity threats, future work will demonstrate how to conduct a similar analysis for privacy.

## Keywords

*Cybersecurity Framework Profile; DNA sequencing; genomics; genomic data; genomic sequencing; human genome; threat modeling; threat mitigations.*

## Feedback

NIST welcomes feedback and input on any aspect of NIST CSWP 35 and additionally proposes a list of non-exhaustive questions and topics for consideration:

1. How well do the threat modeling practices in this white paper relate to existing threat modeling practices leveraged by your organization? Are there significant gaps between the sets of practices that this paper should address?
2. How do you expect this white paper to influence your future practices and processes?
3. How do you envision using this white paper? What changes would you like to see to increase/improve that use?
4. What suggestions do you have on changing the format of the information provided? Would it help to provide a more concise overview document with additional detail provided in either appendices or as part of a more interactive website (e.g., GitHub Pages as used in the NCCoE Zero Trust project)?
5. Is the example provided here sufficient for your organization to identify and address cybersecurity threats in genomic data sequencing or genomic data analysis? Are there changes or additional content that the authors should consider?

## Table of Contents

## List of Tables



## List of Figures

## Executive Summary

In this paper, the National Cybersecurity Center of Excellence (NCCoE) Genomic Data project team demonstrates how to conduct cybersecurity threat modeling against the environments involved in genomic sequencing and analysis. The paper demonstrates a common four-step threat modeling process that can be used as an example for organizations involved in genomic research, sequencing, and analysis planning to conduct similar threat modeling and identify mitigations:

1. Document ***"What are we working on?"*** through architecture, dataflow, and high-value dataflow diagrams for the genomic data processing environment (Sec. 2.1).
2. Evaluate ***"What could go wrong?"*** by identifying threats in the environment using tools such as STRIDE, MITRE ATT&CK®, and attack trees (Sec. 2.2).
3. Determine ***"What are we going to do about it?"*** by prioritizing the identified threats to help sequence and select initial targets for mitigations, leveraging best practice guides and existing resources (Sec. 2.3).
4. Consider ***"Did we do a good job?"*** by reviewing the results of the threat modeling exercise and identifying any additional activities, including high-priority areas where additional mitigations are needed (Sec. 2.4).

**Background.** Legislation such as the Genetic Information Nondiscrimination Act of 2008 (GINA) [1] identifies the need to protect genetic data, while Executive Order 14018 [2] lays out the need to identify risks and develop a protection plan for biological datasets, including genomic data. Cyber attacks may impact the confidentiality, integrity, and availability of systems that process genomic data[1], introducing economic, privacy, discrimination, and national security risks. Organizations rely on genomic data sharing and aggregation to advance scientific and medical research, improve health outcomes, and compete within the global bioeconomy. Cybersecurity and privacy for genomic data are complicated by the nature of the data, which is immutable and includes kinship, health, and phenotype, as well as the broad, diverse, and international composition of the genomics community, which includes government, academia, and industry stakeholders engaged in biopharmaceutical research, healthcare, law enforcement, agriculture, and direct-to-consumer genetic testing.

The paper is part of a larger effort at the NCCoE to engage genomic data processing stakeholders to create practical guidance that addresses related cybersecurity and privacy concerns. The NCCoE Genomic Data website provides links to previous workshops and publications, including National Institute of Standards and Technology (NIST) Internal Report (IR) 8432, *Cybersecurity of Genomic Data* [3], and IR 8467, *Genomic Data Cybersecurity and Privacy Frameworks Community Profile (Genomic Data Profile)* [4]. Additionally, the NCCoE is

[1] Data processing refers to "the complex and interconnected relationships among entities involved in creating or deploying systems, products, or services or any components that process data." NIST Privacy Framework 1.0 https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

181 currently developing a privacy-focused guide to address privacy-related concerns, threats, and
182 risks that will also be published.[2]

[2] While cybersecurity threat modeling can support some privacy needs, additional privacy threat modeling efforts are necessary to address the full scope of privacy. For more information regarding the relationship between cybersecurity and privacy risk management, see the *Genomic Data Profile* [4].

# 1. Introduction

This document provides an example of how to conduct cybersecurity threat modeling on genomic data processing environments to help identify potential cybersecurity threats, their impacts, and potential mitigations. The environments represent a basic implementation with devices, processes, and tools commonly used by government, academia, and industry for processing genomic data.

***Organizations processing genomic data can use the threat modeling techniques and results from this paper to manage cybersecurity threats and reduce cybersecurity risk.***

## 1.1. Use Case and Scope

This threat modeling example addresses the common use case of sequencing deoxyribonucleic acid (DNA) and analyzing the results. The bioeconomy[3] relies on this use case for many of its products and services. The requesting organization (***Research Partner***) sends a physical "wet lab" DNA sample and associated metadata (in digital form) to a ***Genomic Sequencing Laboratory*** that processes the sample and returns the digital results in the form of a genomic sequence. The genomic sequence serves as an input to the Research Partner's data analysis pipelines. Genomics as a scientific field has progressed quickly through open sharing of publicly distributed software. The community benefits greatly by freely sharing this software but should also consistently implement appropriate risk management practices whenever using this software in genomic data analysis pipelines. In this paper, we refer to this untrusted, off-the-shelf, custom, or open-source software (OSS) as "untrusted software." Figure 1 illustrates this use case.

[3] The economic activity derived from biotechnology and biomanufacturing is referred to as the bioeconomy [2].

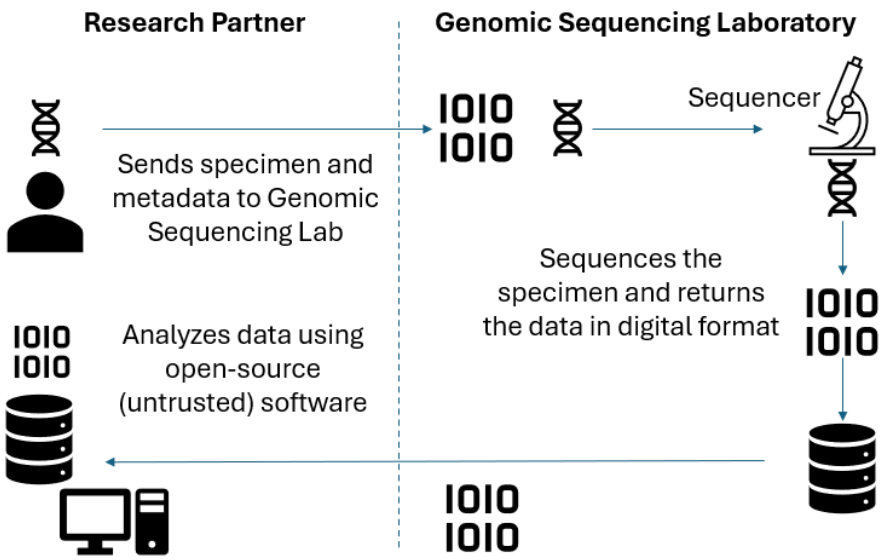



**Figure 1. Genomic Sequencing Workflow**

## 1.2. Organizational Tailoring

Organizations that process genomic data need to protect that data due to both its high value and the privacy risk to individuals if human genomic data are exposed. Organizations need a process to guide the selection of appropriate cybersecurity capabilities to reduce risk to an acceptable level for the confidentiality, integrity, and availability of genomic data. Each organization should consider its own goals and priorities when tailoring this example to select and implement appropriate and cost-effective cybersecurity capabilities to achieve organizational outcomes. The organization should also periodically assess its cybersecurity posture, considering new technologies and threats to identify gaps in cybersecurity outcomes and prioritize mitigations.

NIST IR 8467, the *Genomic Data Profile,* provides a prioritized list of Mission Objectives for organizations processing genomic data and prioritizes CSF 2.0 Subcategories (or outcomes) to support achieving those Mission Objectives. Based on the use case of sequencing genomic material, the project team selected three relevant Mission Objectives from the *Genomic Data Profile* [4], shown in Table 1.

**Table 1. Genomic Sequencing Workflow Mission Objectives**

| Mission Objectives from the Genomic Data Profile | Mission Objective Description (Keyword) |
|---|---|
| **1** | Manage provenance and data quality throughout the genomic data lifecycle (**Data**) |
| **3** | Identify, model, and address cybersecurity and privacy risks of processing genomic data (**Risks**) |
| **8** | Facilitate research and education to advance science and technology (**Research**) |

Throughout the paper, CSF 2.0 Subcategories that were prioritized for one or more of the Mission Objectives in the *Genomic Data Profile* are listed in parentheses and abbreviated as (CSF Subcategory; Mission Objective). For example, the CSF Subcategory GV.OC-01: "The organizational mission is understood and informs cybersecurity risk management" comes from the Govern (GV) Function and the Organizational Context (OC) Category. It received priority designation for Mission Objective 8 and would be abbreviated as (GV.OC-01; MO:8).

## 1.3. Threats and Risks

In the bioeconomy, organizations will have differing Mission Objectives and, therefore, different risks despite facing similar cybersecurity threats. The same threat may have a different impact or likelihood in two different organizations or use cases. For example, a denial-of-service threat may represent a high impact for time-sensitive disease surveillance but a low impact for an agricultural researcher. To maximize the applicability of this paper's use case (sequencing genomic material), the process focuses on threats instead of risks, which are specific to the organization and its use case. The term "threat" is not the same as "risk."

- A threat is "any circumstance or event with the potential to adversely impact organizational operations (including mission, functions, image, or reputation), organizational assets, or individuals" [5][6].
- A risk is "a measure of the extent to which an entity is threatened by a potential circumstance or event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence" [5][7].

Threat modeling scenarios are adaptable to different stakeholders who can bring to the threat model their specific organization-dependent (i) adverse impact and (ii) likelihood of occurrence of the threat that are required to calculate their risk. Threat modeling scenarios can even accommodate different risk and vulnerability assessments beyond that described above as may be appropriate for different use case scenarios (e.g., for the use case of a medical device manufacturer submitting a device for the U.S. Food and Drug Administration (FDA) clearance [8]).

The determination of the potential risk will guide an organization in their risk strategy to eliminate, mitigate, accept, or transfer responsibility for threats to meet their organization's specific risk tolerance and applicable legal or regulatory requirements.

***Organizational risk can be defined as a combination of the likelihood of occurrence of threat becoming realized and the impact that it has on the organization.***

## 1.4. Threat Modeling Overview

***The threat modeling process identifies cybersecurity objectives and vulnerabilities across the system and defines countermeasures to eliminate, mitigate, accept, or transfer responsibility for threats throughout the system's lifecycle.***

The NCCoE team used the Four Question Framework [7], illustrated in Figure 2, to structure the threat modeling process by answering:

1) What are we working on?
2) What could go wrong?
3) What are we going to do about it?
4) Did we do a good job?

Though the questions are listed in sequential order, the process is iterative, as shown by the arrows in the figure. Each question is addressed through specific techniques outlined in this paper. Answers to one question may be used to modify previous answers or highlight the incompleteness of an answer to a previous question.

Since some genomic sequencers are regulated as medical devices when used as *in vitro* diagnostic products as defined in 21 CFR Part 809.3, this paper uses the threat modeling approach described in the *Playbook for Threat Modeling Medical Devices (Playbook)* [9] that is based on methods described in the "Threat Modeling Manifesto" [10]. The FDA, in its premarket guidance for cybersecurity in medical devices [8], refers to a threat modeling methodology and recommends that medical device manufacturers implement threat modeling to analyze and identify security concerns in medical devices. The Playbook can also be used as a guide to conduct threat modeling by organizations who are not medical device manufacturers, as is the case in this paper.

Because this threat model is intended for various stakeholders who have differing risks to the same threats, some possible mitigations will be suggested. However, organizations will choose specific mitigations depending on their mission, their goals in performing the threat modeling process, the risks associated with the specific use case, the phase of the system lifecycle, and the resources at their disposal. Therefore, a comprehensive list of mitigations and answers to Question 4 will not be provided in this paper. Section 2.3 provides potential mitigations reflective of common threats and implementations seen in the genomic workflow. Individual organizations can translate threats into risks by incorporating mission and use case-specific probabilities and impacts. The calculated risks can help them choose whether to mitigate, accept, transfer, or eliminate the specific threats.

### 1.5. Audience

This paper is intended for organizations that process genomic datasets. Organizations that sequence genomic material, analyze genomic datasets, or transfer genomic data files can apply a similar threat modeling process, including the sample architecture diagrams, threats identified, suggested mitigations, and other findings from this paper to help them identify and address similar threats in their environments.

## 2. Threat Modeling Example

This section applies the *Playbook* methodology [9] to provide an example of how to conduct threat modeling on the genomic sequencing workflow using the Four Question Framework (Figure 2). For completeness, we include Tables 2, 3, and 4 from the *Playbook* [9] to guide the interpretation of system architecture diagrams that describe the sequencing workflow. We refer readers to the *Playbook* and the Shostack website[4] for more threat modeling examples and thorough descriptions that extend beyond genomics.

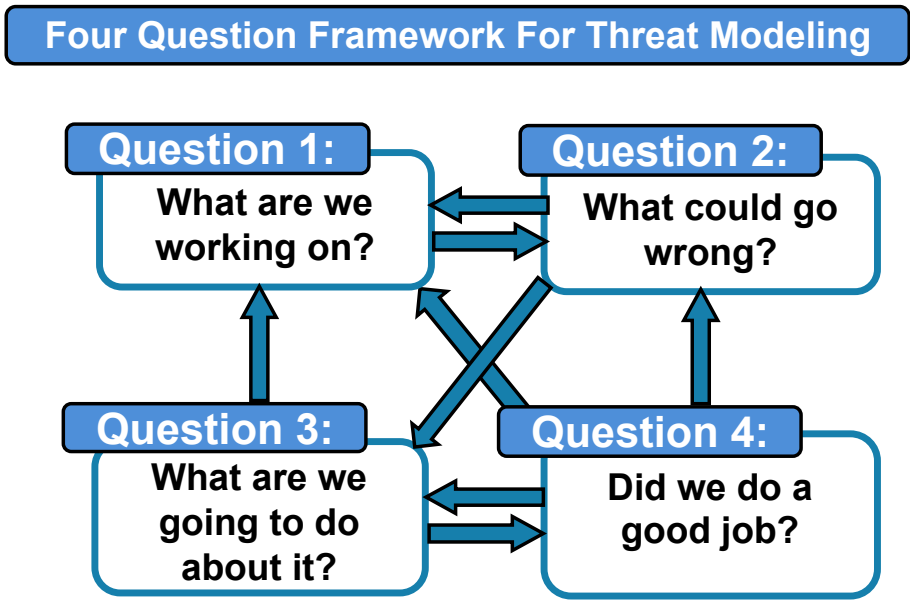



**Figure 2. Visualization of the Four Question Framework for Threat Modeling [11]**

### 2.1. Question 1: What are we working on?

Answering Question 1 helps teams identify activities and language to better understand and describe the system(s) being analyzed. This involves reviewing the system, interviewing associated personnel, analyzing architecture documents, and building out the use case to develop a shared understanding of the system components, functionality, and interfaces. Through this process, the team establishes a baseline understanding that will support analyzing cybersecurity threats against the system, evaluating the effectiveness of cybersecurity mitigations, and characterizing the resilience of the system.

This section identifies and characterizes the system and data of interest using Dataflow Diagrams (DFDs) and High-Value Dataflows (HVDs). First, we describe the diagraming techniques and then apply those techniques to provide example diagrams for both the Genomic Sequencing Laboratory and Research Partner environments.

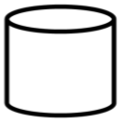
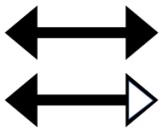
**Dataflow Diagrams.** The team developed DFDs to document "What are we working on?" The DFDs depict trust boundaries and communication paths between different components of the

[4] https://shostack.org

system being analyzed. This technique was selected due to the “system of systems” nature of the use case, since DFDs highlight interactions among external entities and trusted components. DFDs also facilitate the Spoofing, Tampering, Repudiation, Information Disclosure, and Escalation of Privileges (STRIDE) threat analysis, a technique that will be described under Question 2. DFDs help teams produce a common architecture document that can be used for other collaboration and development activities outside the threat modeling effort.

The format of these dataflow diagrams follows conventions established in the *Playbook* and repeated here in Table 2.

**Table 2. Symbols Used in Detailed DFDs**

| Element | Symbol | Discussion |
|---|---|---|
| **External Entity** |  | **Object:** A sharp-cornered rectangle.<br>**Represents:** Anything outside your control. Examples include people and systems run by other organizations or even divisions. |
| **Process** |  | **Object:** A rounded rectangle.<br>**Represents:** Any running code, including compiled, scripts, shell commands, Structured Query Language (SQL) stored procedures, et cetera. |
| **Data Store** |  | **Object:** A drum.<br>**Represents:** Anywhere data are stored, including files, databases, shared memory, cloud storage services, cookies, et cetera. |
| **Dataflows** |  | **Object:** A double-headed arrow.<br>**Represents:** All the ways that processes can talk to data stores or each other. If a conversation is only initiated by one side, you can represent the initiating side as an empty arrow. |
| **Trust Boundary** |  | **Object:** A closed shape drawn with a dashed or dotted line.<br>**Represents:** A way to display different trust levels between objects. |

Each rectangle with dotted lines represents a ***trust boundary***. Each rectangle with solid lines represents a ***component***. All stick figures represent ***human actors*** in the environment. All lines connecting components or actors represent ***dataflows*** that can be either digital or physical (such as a network connection or a human inserting a physical sample into the sequencer). Dataflows that were determined to be HVDs are labeled with a "**D**" followed by a number and are shown in a ***darker line*** than other dataflows. Dataflows are shown as double-headed arrows. A ***hollow arrow*** on one side of a given dataflow implies that the component or process on that side of the dataflow is the initiator of the communication.

**High Value Dataflows.** Some DFDs, called HVDs, were selected for more detailed analysis because they significantly impact the system's security and resiliency, as described in Section 2.1.3. While DFDs help identify which components and processes share data, they do not capture the details of how protocols and organizational use cases operate. To get to that level of detail for HVDs, this paper leveraged cross-functional swim lane diagrams.

Figure 1 illustrates interactions between the Research Partner and the Genomic Sequencing Laboratory to transfer, sequence, and analyze genomic data. The use case can be applied to interactions with other external entities that may include equipment manufacturers, untrusted software, and cloud providers. The following sections provide examples of DFDs and HVDs for both the Genomic Sequencing Laboratory and the Research Partner.

### 2.1.1. Genomics Sequencing Laboratory Data Flow Diagrams

The ***Genomic Sequencing Laboratory*** consists of multiple environments and boundaries that work together to process, analyze, and transfer genomic data. The systems involved in the transfer can be physical or virtual, ranging from laboratory equipment and genetic sequencers to virtualized applications or cloud storage and services.

Figure 3 illustrates the high-level architecture of the Genomic Sequencing Laboratory's boundaries, environments, and systems, using an identifier (ID) to reference each component. The Genomic Sequencing Laboratory environments each have their own trust boundary.

Example environments include a Wet Lab with sequencing (ID: C1), a Management and Tooling environment (ID: C3), a Research and Computing environment (ID: C10), and a Data Delivery Demilitarized Zone (DMZ) environment (ID: C21) for controlling access to the storage environment from outside entities. These separate but integrated environments process physical DNA samples that become genomic data in the form of raw data, metadata, intermediate processed data, and reports. Genomic data can be transferred between environments and across trust boundaries.

The Remote Access and External Actors trust boundary can be found at the top of the diagram. This trust boundary includes all entities external to the Genomic Sequencing Laboratory network that are anticipated to connect to the lab's network. Common examples of external entities include the National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI), as well as Software as a Service (SaaS) or Platform as a Service (PaaS) applications used by the lab.

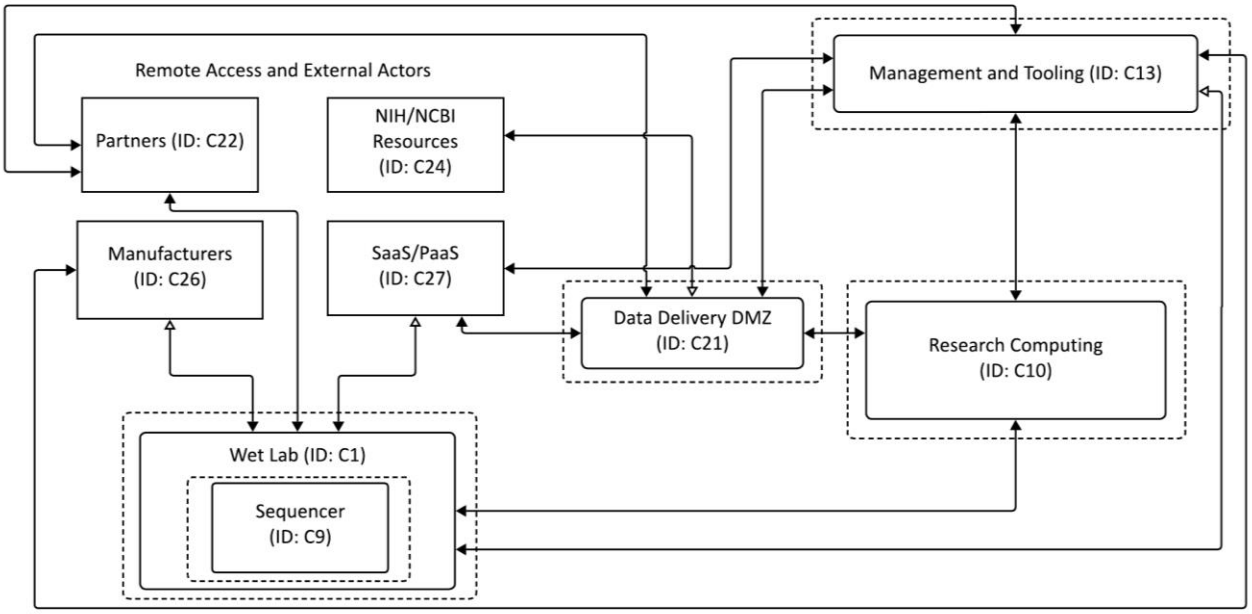
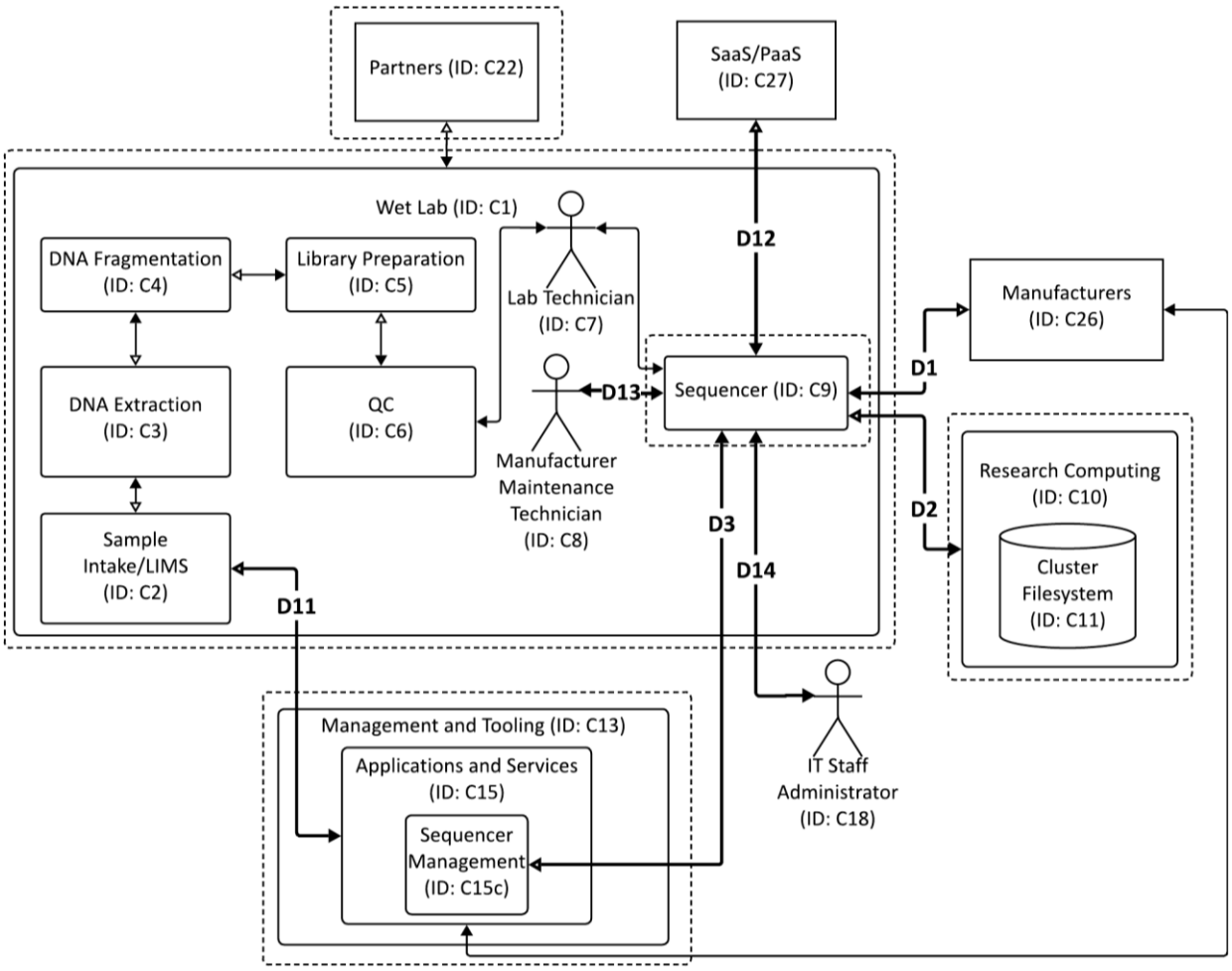



**Figure 3. High-Level Architecture of the Genomics Sequencing Laboratory**

Because of its complexity, more detailed DFDs of the Genomics Laboratory are presented separately for readability and clarity (ID.AM-03; MO:1,8) in Figures 4 through 7.

Figure 4 represents a more detailed DFD of the ***Wet Lab*** (ID: C1) and its associated process. The Wet Lab contains the equipment necessary to analyze physical DNA samples, digitize the genomic information, and combine the data with the Laboratory Information Management System (LIMS) digital data (ID: C2) that identifies the physical sample. This includes equipment used during the DNA Extraction (ID: C3), DNA Fragmentation (ID: C4), Library Preparation (ID: C5), Quality Control (ID: C6), and sequencing phases of the genomic data lifecycle. The flows of data between these components of the Wet Lab are shown as connections between components on the diagram. The Lab Technician (ID: C7) and the Manufacturer Maintenance Technician (ID: C8) are shown as well, as they interact directly with Wet Lab systems and are within the Wet Lab trust boundaries.




371 **Figure 4. Wet Lab DFD**

Figure 5. Sequencer DFD shows the ***Sequencer*** (ID: C9) with much more detail regarding internal logical components in comparison with other Wet Lab equipment. We highlighted the Sequencer (ID: C9) as a device of interest to the threat modeling effort due to the complexity, high cost, and comparatively large threat surface within the Wet Lab. The Sequencer is the device in the Wet Lab network that converts a physical sample into digital DNA sequence data. The Sequencer allows direct connections from the Manufacturers (ID: C26) at all times for the purposes of remote maintenance. This always-available connection introduces potential security threats.

Once the DNA sequence data leaves the Wet Lab, it travels to the Cluster Filesystem (ID: C11) within the Research Computing (ID: C10) environment, where the data are then stored.

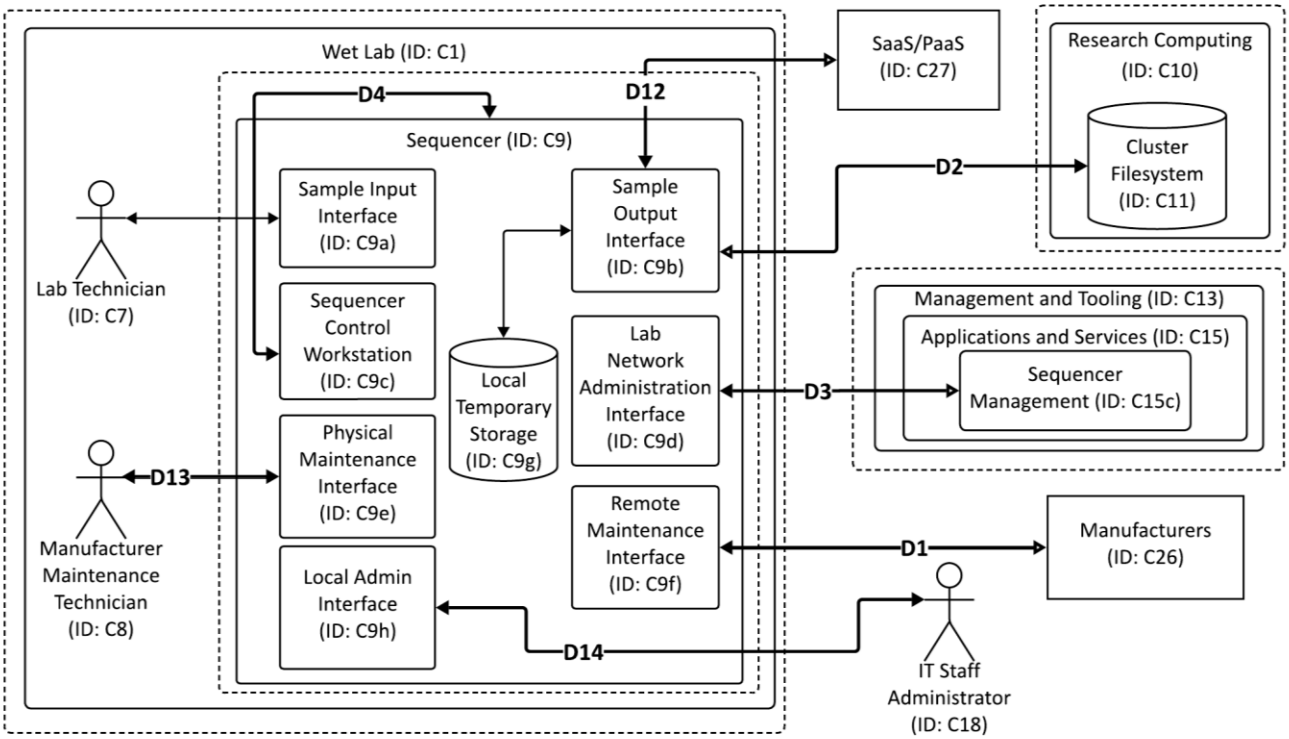



**Figure 5. Sequencer DFD**

Figure 6 provides more details for the Data Delivery DMZ (ID: C21), Management and Tooling (ID: C13), and Research Computing (ID: C10) environments. Data are delivered from the Sequencer to a restricted area on the Cluster Filesystem (ID: C11) within the Research Computing Environment. The Quality Control (QC) analysis (ID: C12) can perform operations on the data within this restricted area on the Cluster Filesystem. When ready, the data are copied from the restricted location on the Cluster Filesystem to an area on the Cluster Filesystem that can be accessed by the Data Delivery DMZ (ID: C21) for delivery to an external entity.

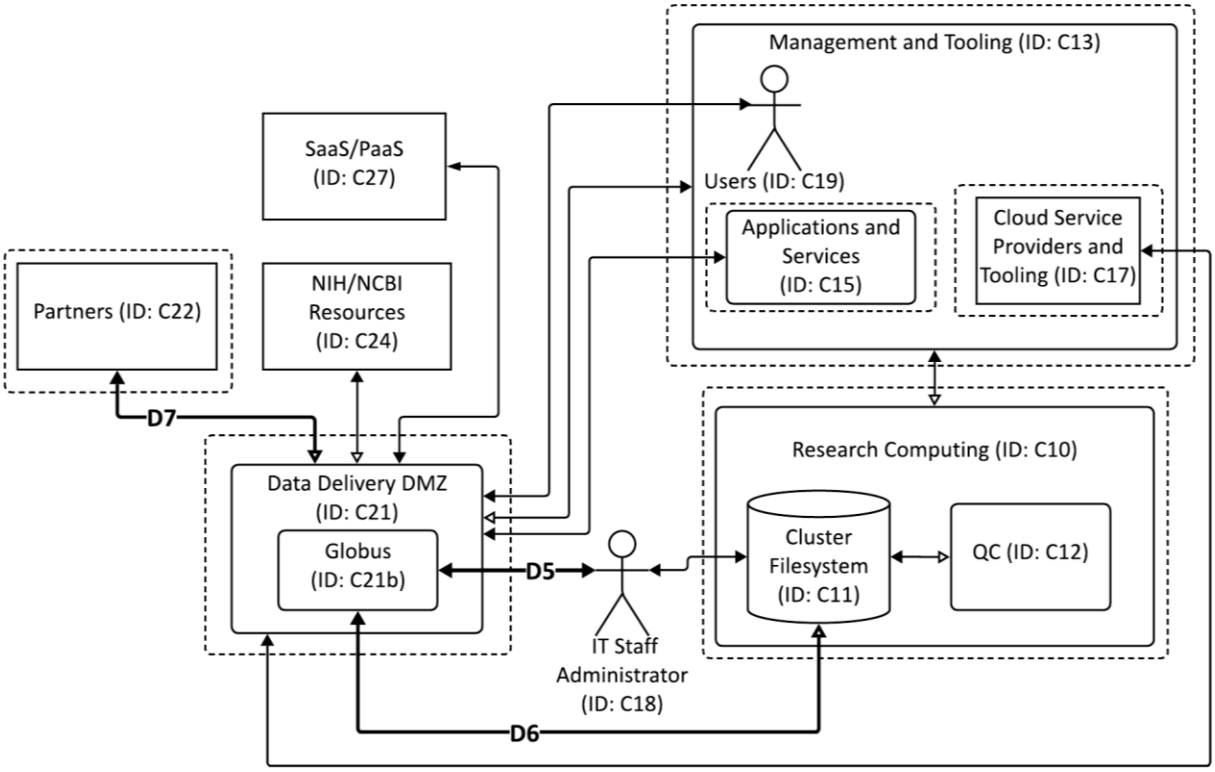



**Figure 6. Data Transfer DFD**

Figure 7 depicts the environment that houses Monitoring and Security Logs (ID: C14b), Cyber Tooling (ID: C15d), Administration (Admin) and Information Technology (IT) or Admin/IT Tooling (ID: C15b), and Sequencer Management (ID: C15c). In one embodiment, these would be virtual machines (VMs) running on a server that has a Hypervisor (ID: C16).

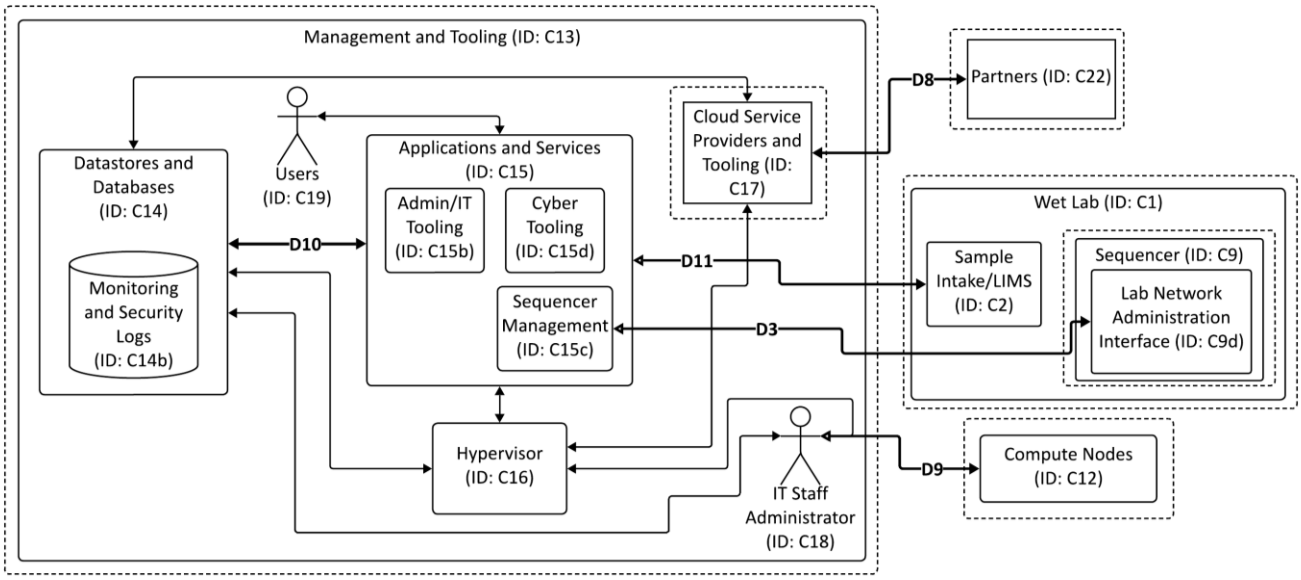



**Figure 7. Management and Tooling DFD**

### 2.1.2. Research Partner Data Flow Diagrams

Figure 8 illustrates a high-level architecture of the **Research Partner** environment. The main system of interest for the Research Partner is the Compute Server (ID: 101). Threats against the confidentiality, integrity, and availability of this system are of particular importance to the Research Partner threat model, as this is the system where bioinformatics analysis takes place using sensitive genomic data that is stored on the server.

The Research Partner environment connects to external entities, including the Genomic Sequencing Laboratory (ID: 108) to sequence the data, Manufacturer Updates (ID: 127) for updates to the operating system (OS), Untrusted Software Packages (ID: 104) used for genomic analysis, and Genomic Reference Resources (ID: 105) required for genomic analysis.

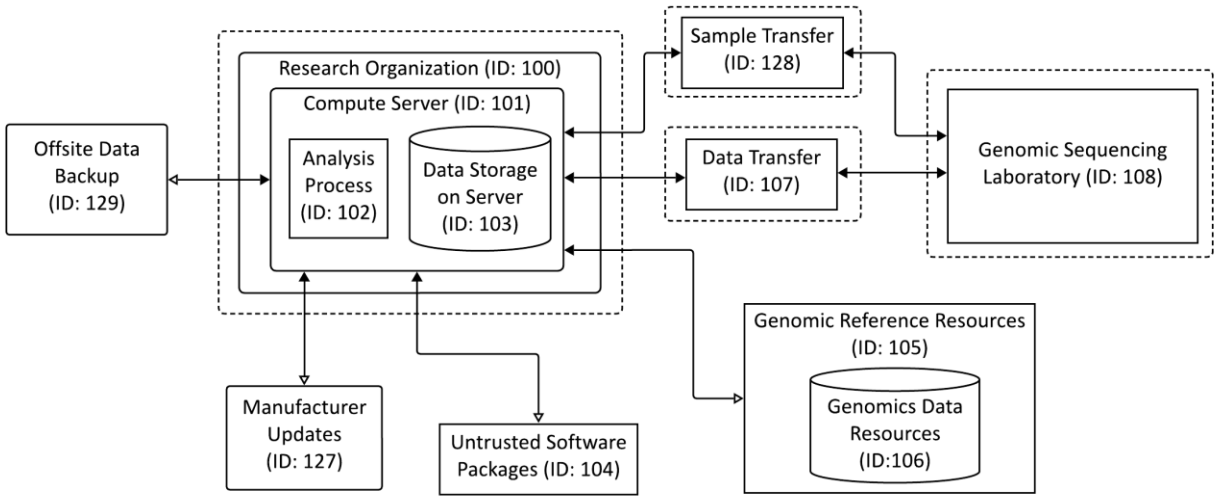



**Figure 8. High Level DFD for Research Partner**

Figure 9 illustrates a more detailed architecture of the Research Partner environment with associated personnel and external connections (ID.AM-03; MO:1,8). An administrator who performs OS and IT Setup (ID: 130) manages the server and endpoint protection of the server. A Bioinformatics Analyst (ID: 131) uses the server for genomic analysis and may initiate transfers of data or software between trust boundaries.

Each of the external entities identified introduces potential threats to be considered from the perspective of the Research Partner environment (ID.AM-04; MO:1,8). Some of the external entities are likely more trusted (such as NIH/NCBI or Globus[5]) than others. Some OSS projects are security conscious and follow best practices such as the recommendation of the Secure Software Development Framework (SSDF) [12], including inviting the public to review the code for security vulnerabilities and submit improvements. However, some OSS may be considered less trusted if it does not follow SSDF recommendations, resides in publicly accessible repositories without secure access and change control, or has maintenance that is heterogenous depending on the career trajectory of the researchers that initially develop the software.

Dataflow D102 in Figure 9 connects the Globus Transfer Client (ID: 123) with the Data Delivery DMZ (ID: 126), also labeled as ID: C21 in Figure 6. This represents that the Globus Server Connect application is running in the Data Delivery DMZ (ID: 126) of the Genomic Sequencing Laboratory, and the Globus Personal Connect Client is running on the Server (ID: 101).

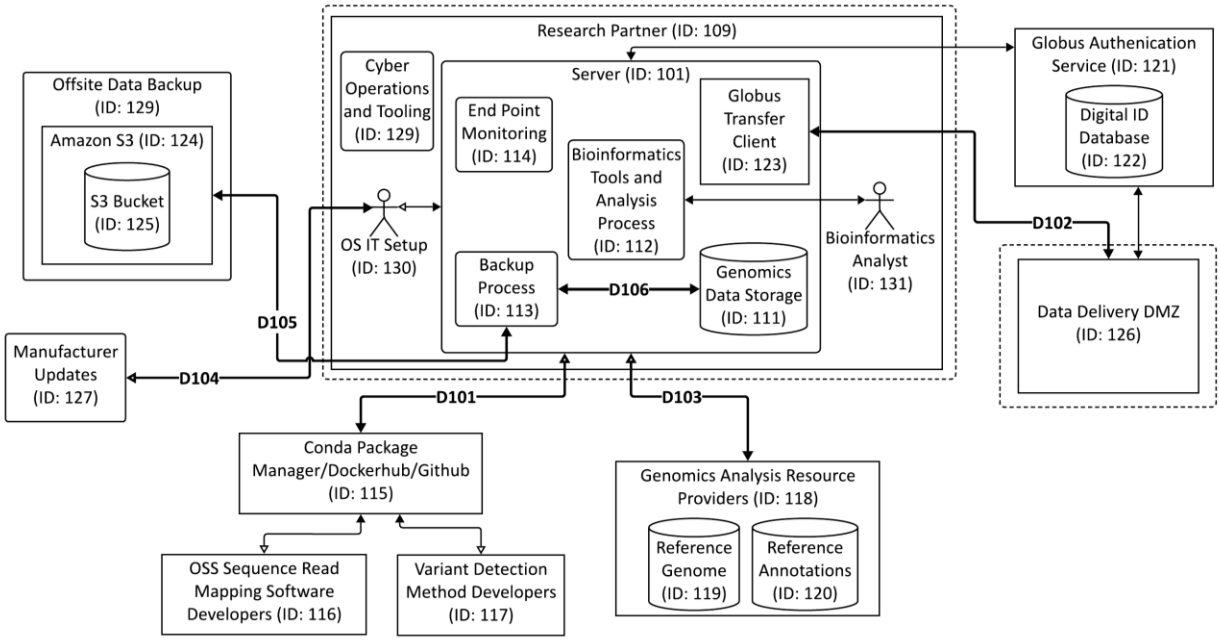



**Figure 9. Detailed DFD of Research Partner**

[5] Globus is research cyberinfrastructure for securely moving, sharing, and discovering data, developed and operated as a nonprofit service by the University of Chicago: https://www.globus.org/what-we-do. For technical details, see https://docs.globus.org/guides/recipes/modern-research-data-portal/.

### 2.1.3. High-Value Dataflows Overview

DFDs are useful for depicting which components communicate with each other, but they are static models that do not capture the details of how protocols and use cases operate. DFDs can be used to identify HVDs that merit detailed analysis using different modeling techniques. This section describes the use of HVDs as a modeling technique and identifies six example HVDs, three from the Genomic Sequencing Laboratory and three from the Research Partner (ID.RA-05; MO:1,3,8). Comprehensive threat modeling, as may be needed to comply with regulatory requirements, will address all HVDs in a similar way to the six specific examples in this paper.

In our documentation, HVDs are processes or use cases based on areas of interest, often because they were highlighted in the *Playbook,* cross trust boundaries, perform a critical function, or access a critical system. HVDs tend to be high risk and have a high impact if they are compromised.

***DFDs can be designated as HVDs when they cross trust boundaries, perform a critical function, or access a critical system.***

Examples of HVDs from the *Playbook* that are relevant to this system include:

- Authentication protocols
- Programming and configuration commands
- Obtaining and validating software updates
- Procedures to restore from backups

Modeling techniques of system state can be helpful to describe how the system will be used, the different modes it may find itself in, and how the system handles error states and invalid input. For example, a genomic sequencer may behave differently if it is in a sequencing mode versus a service mode. One modeling technique that can be useful is a cross-functional swim lane diagram. This paper leveraged cross-functional swim lane diagrams to document the details of processes and use cases that were identified as HVDs, potentially having a large impact on the security or resiliency of the system.

The selection of HVDs was guided by the following considerations:

- The *Playbook* has a brief list of HVDs for medical devices. When used for clinical diagnostics, a high throughput genomic sequence analyzer is regulated as a Class 2 medical device [13], so all HVDs listed in the *Playbook* were carefully considered.
- Dataflows that cross multiple trust boundaries also received consideration to be selected as an HVD, as traditionally these types of dataflows have large attack surfaces and are often entry points for adversaries.
- The most valuable assets in both the Genomic Sequencing Laboratory and the Research Partner were the genomic data. In the Genomic Sequencing Laboratory environment, the data reside on the Cluster Filesystem (ID: C11), and in the Research Partner environment, the Data Storage on Server (ID: 103) (ID.AM-05; MO:1,3). Thus, dataflows that interacted with either of these highest-value data were identified as key HVDs.

### 2.1.4. Genomic Sequencing Laboratory HVD Examples

The diagrams that follow are a subset of the identified HVDs from Figures 4 through 7. These examples provide more detailed cross-functional diagrams with accompanying text to demonstrate the detail needed to answer Question 1 (“What are we working on?”) in sufficient detail so that it can provide the required input to Question 2 (“What can go wrong?”) of the threat modeling process.

#### 2.1.4.1. Example HVD 1

The first important HVD considered in the Genomics Sequencing Laboratory architecture is the connection between the Cluster Filesystem (ID: C11) within the Research Computing Environment and the Data Delivery DMZ (ID: C21) for controlling access to the storage environment from outside entities. This dataflow is labeled as D6 in Figure 6 DFD. Because it hosts the genomic data sequences, the Cluster Filesystem is one of the most valuable assets of the Genomic Sequencing Laboratory environment. The Data Delivery DMZ is important to consider because it is exposed to the hostile internet.

Figure 10 diagrams the data transfer process where an administrator in charge of fulfilling data requests for Research Partners requests a list of available files from the Cluster Filesystem. The Cluster Filesystem provides the files after it checks whether the administrator has the necessary permissions to read and publish the files. Then the administrator selects which files should be copied to the Data Delivery DMZ-accessible folder, and the Cluster Filesystem copies the files into an area for Research Partners via the Data Delivery DMZ.

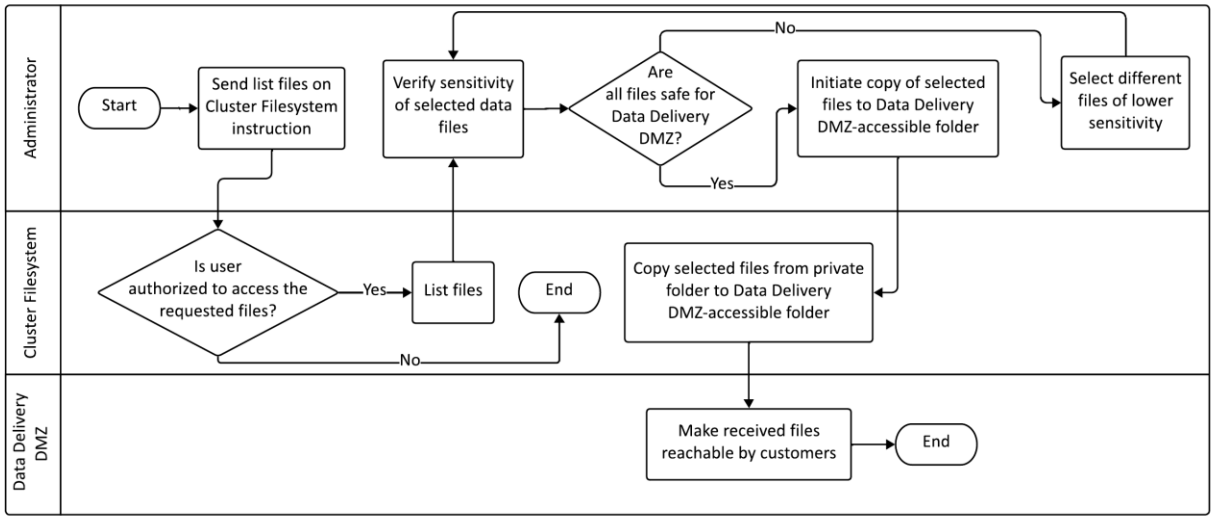



**Figure 10. HVD 1: Cluster Filesystem to Data Delivery DMZ**

### 2.1.4.2. Example HVD 2

The second important HVD considered is labeled D1 in Figure 5, the connection between the Manufacturer (ID: C26) and the Remote Maintenance Interface (ID: C9f) of the sequencer. This connection is used, among other things, for sequencer software updates. This is considered an HVD because the entire use case relies on the integrity and data quality of the sequencing. The Cluster Filesystem within the Research Computing Environment also trusts the integrity and data quality of the sequencer. Additionally, the connection to the manufacturer happens via the untrusted internet, and some manufacturers and/or their service departments may be in a country or location of concern for the organization. Organizations should evaluate the risks from equipment, software, and processing that involve locations of concern.

Figure 11 illustrates the cross-functional diagram as an example of a process that a manufacturer might use in updating the sequencer software. The manufacturer first connects to the sequencer and provide the necessary credentials for the sequencer to authenticate. After successful authentication, the manufacturer sends the updated binary file and installs it on the sequencer. The session concludes with the manufacturer verifying that the update was successful and that the sequencer is functioning properly.

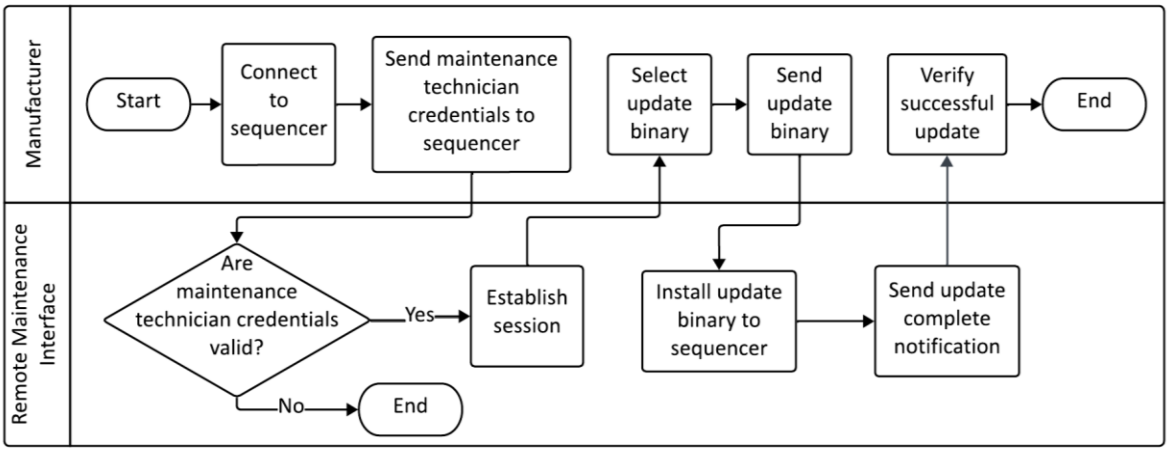



**Figure 11. HVD 2: Sequencer Remote Interface to Manufacturer**

### 2.1.4.3. Example HVD 3

The connection between the Sample Output Interface (ID: C9b) from the Wet Lab sequencer and the Cluster Filesystem (ID: C11) [labeled as D2 in Figure 5] within the Research Computing Environment is the third HVD diagramed. This dataflow connects the two most valuable components of the environment with the data flowing across a trust boundary. Depending on the configuration, the sequencer may be trusted by the Cluster Filesystem and vice versa.

Figure 12 illustrates the cross-functional diagram for this HVD in more detail. The sequencer sequences a sample. When finished, it checks whether the results have any errors. If not, it sends the data to the Cluster Filesystem when a network connection exists with the Wet Lab sequencer. If there is no network connection, the sample output interface temporarily saves the data to secondary storage. The sequencer may be trusted by the Cluster Filesystem.

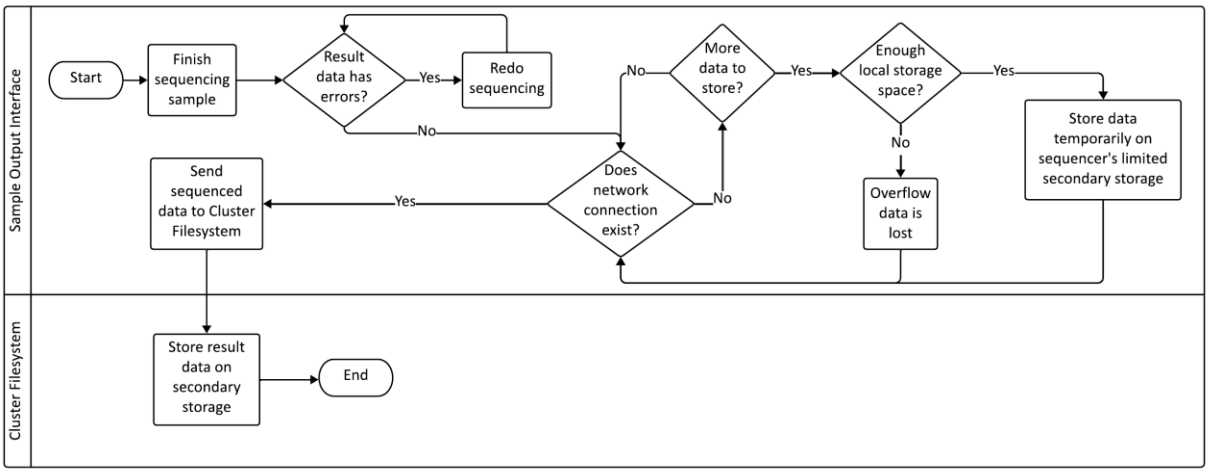



**Figure 12. HVD 3: Sequencer to Cluster File System**

Other HVDs in Figures 4 through 7 could be included in the threat modeling but were not described as one of the three examples of Genomic Sequencing Laboratory HVDs. Sorted by figure, these include:

- **Wet Lab** (Figure 4) HVD D11: The Sample Intake (ID: C2) is where the lab receives the highly valuable physical genomic sample and shares the sample metadata with Applications and Services (ID: C15c).
- **Sequencer** (Figure 5) HVD D3, D4, D12, D13, D14: The sequencer dataflows crossing trust boundaries include HVD D3—remote Sequencer Management (ID: C15c) and HVD D12—the connection between the Sample Output Interface (ID: C9b) and SaaS/PaaS (ID: C27) storage. HVD D4 identifies the user interface to the Sequencer (ID: C9) from the Sequencer Control Workstation (ID: C9c), which may be either a separate workstation or integrated with the sequencer. HVD D13 and D14 capture software updates to the sequencer, including those made by the Manufacturer Maintenance Technician (ID: C8) and IT Staff Administrator (ID: C18).
- **Data Delivery DMZ** (Figure 6) HVD D5, D6, and D7: These connections to the Data Delivery DMZ (ID: C21) include IT Staff Administrator (ID: C18) who initiate data transfers in Globus, storage to the Cluster Filesystem (ID: C11), and dataflows with Partners (ID: C22).
- **Management and Tooling** (Figure 7**Error! Reference source not found.**) HVD D8, D9, a nd D10: These connections to the Management and Tooling environment (ID: C13) include HVD D8—the external connection to third-party Partners (ID: C22), HVD D9—the remote administrative access of the Compute Nodes (ID: C12), and HVD D10—application and services connections to confidential internal data (ID: C14).

### 2.1.5. Research Partner HVDs

This section provides examples of HVDs and cross-functional diagrams for the Research Partner environments. The location of the HVDs in the environment is shown by bold connections labeled D101 through D105 in Figure 9.

#### 2.1.5.1. Example HVD 4

The fourth example of an HVD (D101 in Figure 9) in a bioinformatics analysis environment is the identification, installation, and use of untrusted research software by the Research Partner. Many research conferences and articles detail newly available software, along with GitHub links or other download mechanisms. This software then runs directly on the high-value genome sequencing data. Figure 13 illustrates the cross-functional diagram for this process.

The researcher identifies the new research software to use and the location where it is stored. The researcher then identifies an appropriate configuration of access rights granted through a policy enforcement module (e.g., as referenced in Figure 13 by the use of an AppArmor or SELinux profile) that limits the software's privileges and processes to only those that are appropriate to perform its function.

After connecting to the server with Secure Shell (SSH), the researcher either uses the appropriate profile identified to create a container for the software or develops a new profile by logging requests that the application will use. After creating the container with the appropriate profile, the software is installed in the container and access is given to the container with the genomic data that the researcher wants to process with the new research software. As an added precaution, the software distribution system performs software testing that includes scanning for malware prior to use.

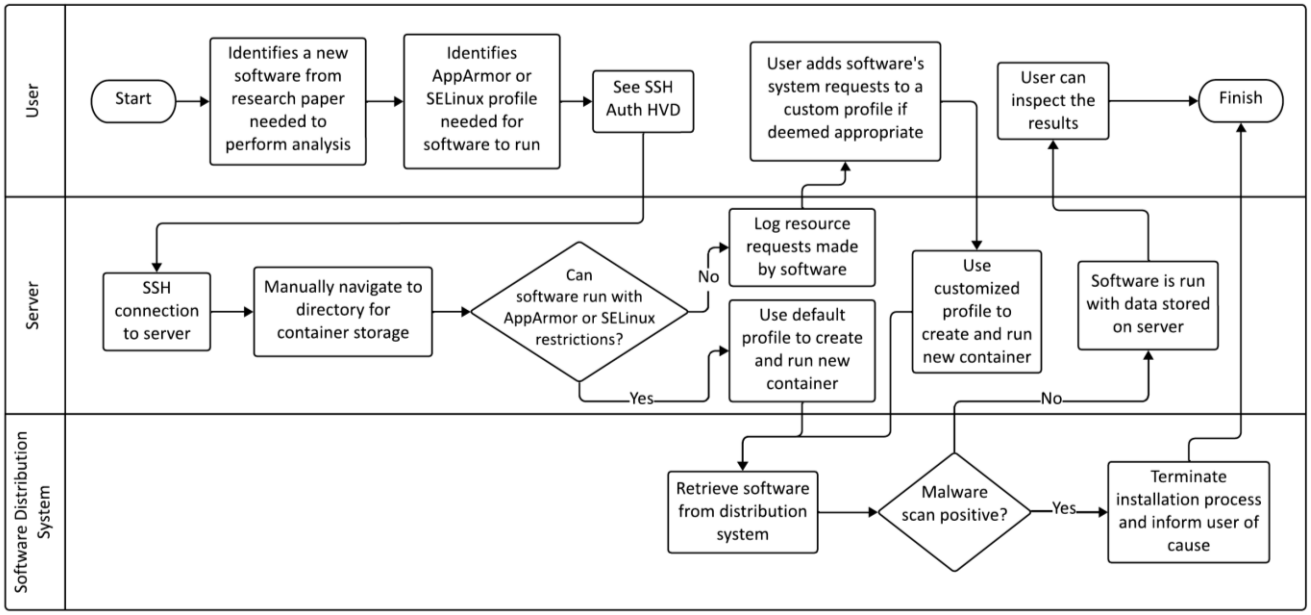



**Figure 13. HVD 4: Example of a Method of Running Untrusted Software on Genomic Data**

### 2.1.5.2. Example HVD 5

Figure 14 identifies the fifth HVD for consideration, labeled as D105 in Figure 9. This HVD includes the processes used to back up sequencing data files for recovery after an equipment failure or from a ransomware attack. The genomic data are of great value to the Research Partner, representing a significant cost to replace. However, the genomic datasets can be very large, so typical enterprise backup solutions may not suffice.

As shown in Figure 14, a user starts the process by receiving a notification from a scheduled backup (the “cron job”) that genomic data in their local storage is out of sync with backup data. This cron job alerts a user that no backup has taken place or that these data are not properly accounted for in a backup. The user will then connect to the server to determine if the data that needs to be backed up are encrypted. If encrypted, the user will transfer the encrypted data to the backup service. If the data are not, the user will need to encrypt the data. That data encryption will also represent an HVD because of the value of protecting the encryption keys.

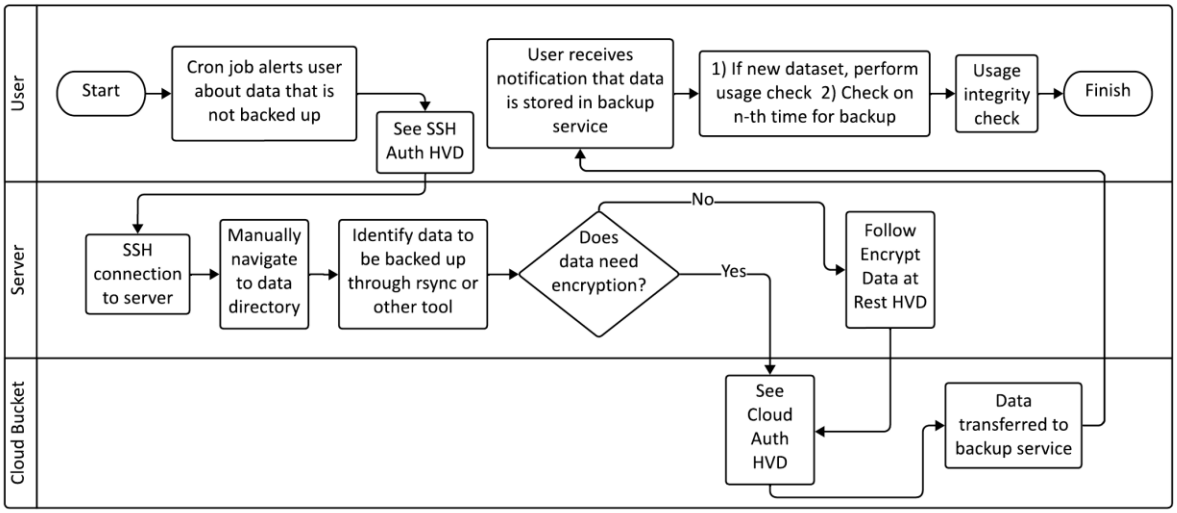



**Figure 14. HVD 5: Backing up Sequencing Data to Cloud Storage**

### 2.1.5.3. Example HVD 6

Figure 15 illustrates the sixth HVD, the receipt of genomic sequencing data from a Research Partner, labeled as D102 in Figure 9. Securely transferring data between a Research Partner and a Genomic Sequencing Laboratory environment, in this case, is mediated by the Globus research transfer service. Figure 15 illustrates this scenario where a user receives a communication, such as an email from the Research Partner, that data are ready for transfer from an agreed-upon transfer endpoint. The user authenticates with Globus and initiates the data transfer to the server using the Globus GridFTP user interface (UI) with the "encrypt transfer" option selected. Upon receiving the Globus transfer completion notification, the user verifies the signature of the file (which mitigates a tamper threat in Section 2.2.2 described below) to ensure the integrity of the data (PR.DS-01: MO:1,3,8). The data can then be used for analysis.

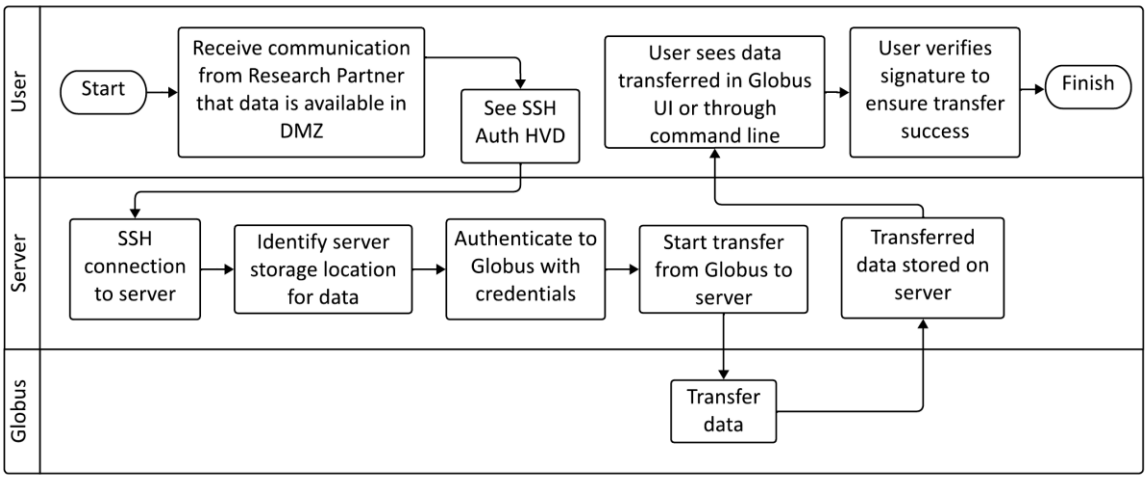



**Figure 15. HVD 6: Obtaining Genomic Data from Genomic Sequencing Laboratory**

Other HVD flows identified in Figure 9 but not described as one of the six HVDs above include:

- HVD D103: The dataflow between the Research Partner (ID: 109) and publicly available Genomic Analysis Resource Providers (such as the NIH/NCBI) (ID: 118) that contain reference genomes and annotations.
- HVD D104: Updates to the Research Partner Server (ID: 110) OS through the OS IT Setup (ID: 130).
- HVD D106: Storing and encrypting the genomic data at rest, either on local Genomic Data Storage (ID: 111) or for Backups (ID: 113). Here the organization should decide whether to encrypt and how users will manage keys to decrypt.

## 2.2. Question 2: What could go wrong?

After DFDs were prepared, HVDs were identified, and the question "What are we working on?" seemed to be adequately addressed, work began on the second question, "What could go wrong?" While the *Playbook* details several techniques for identifying threats, the team used two methodologies, STRIDE and attack trees based on MITRE ATT&CK Tactics, Techniques, and Procedures (TTPs). Note that as the team worked on question 2, it sometimes revealed that question 1 needed additional details added for completeness.

The STRIDE and MITRE ATT&CK methodologies were supplemented by determining priorities from the stakeholders that identified resources that were the most important to protect. For the Research Partner, the genomic datastore was prioritized as the most valuable element to protect. For the Genomics Sequencing Laboratory, the genomic datastore was also considered the most important asset, followed closely by the genomic sequencer (some stakeholders might reverse the importance of the two). The genomic sequencer is expensive and a key part of the revenue generation of commercial laboratories. These identified high-value assets were used to prioritize the key STRIDE threats and to identify the assets targeted by the attack trees.

### 2.2.1. Spoofing, Tampering, Repudiation, Information Disclosure, and Elevation of Privilege (STRIDE)

The STRIDE methodology involves identifying and organizing threats from these six STRIDE elements (or categories) against individual components of the system being analyzed. Sometimes threats overlap categories. For example, ransomware that encrypts data and requires payment for the encryption key could be classified as either tampering or denial of service. During this exercise, capturing the threat is more important than classifying it as one type of STRIDE element or another.

The STRIDE methodology does not rely upon analyzing past attacks and disclosures. This makes STRIDE well-suited for understanding potential future threats for newly developed systems and capabilities. Table 3 describes the STRIDE elements and provides genomic examples for each element.

**Table 3. STRIDE Mnemonic with Examples**[6]

| STRIDE Element | Description | Example |
|---|---|---|
| **Spoofing** | Tricking a system into believing a false entity is a true entity | Using stolen or borrowed credentials to log on as an authorized researcher |
| **Tampering** | Intentional modification of a system or data in an unauthorized manner | Modifying genomic data to stealthily add pathogenicity |
| **Repudiation** | Disputing the authenticity of an action taken | Denying that you accessed other researchers' genomic data |
| **Information Disclosure** | Exposing information intended to have restricted access levels | Publishing a Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) guide Ribonucleic Acid (RNA) sequence that is a trade secret in commercial development |
| **Denial of Service (DoS)** | Blocking legitimate access to the functionality of a system by malicious process(es) | Sending a Transmission Control Protocol (TCP) packet flood to prevent genomic data transfer between systems on the internet |
| **Elevation of Privilege (EoP)** | Gaining access to functions to which an attacker should not normally have access according to the intended security policy | A researcher using a vulnerability in a genomic data transfer web portal to access other researchers' genomic data, rather than just their own |

STRIDE has the advantage of being very structured and can improve brainstorming by ensuring DFD elements (such as processes, datastores, dataflows, and external entities) are not ignored. It can be used by threat modelers of all experience levels to identify threats to the system independent of selecting effective mitigations. However, often there are costs to mitigations, and STRIDE's weakness is that it fails to tell a story of how a threat might represent a real risk. For example, the threat may be difficult to exploit because of other mitigations that an attacker would need to bypass to get to the process or dataflow that the threat is against. The STRIDE methodology does not inform prioritization of mitigations and justifying cost or risk trade-offs. The *Genomic Data Profile* [4] or attack trees can be used to prioritize mitigations.

The STRIDE analysis was performed for the detailed DFDs of the Genomic Sequencing Laboratory and the Research Partner. The team iterated the analysis as the understanding and models were refined.

Each component was given a unique component identifier (Component ID). Every threat against a component was given a unique identifier as well (Threat ID) and classified as one of the six STRIDE threats. No attempt was made to sequentially or otherwise assign the Component or Threat IDs, except to keep them unique. Figure 16 illustrates the format of the table elements for the STRIDE analysis. Each row includes the component and unique threats identified against that element.

[6] Reproduced from the *Playbook* [9] and modified with genomics examples.

| Component Name/ID | Spoof | Tamper | Repudiate | Info Disclosure | DoS | EoP |
|---|---|---|---|---|---|---|
| Remote Maintenance Interface (ID: 9f) | | 47 | 48 | | 49a, 49b | 50 |
| Local Temporary Datastore (ID: 9g) | | 51 | | 52 | 53 | |
| Data Delivery DMZ (ID: 21) | 54 | 55 | 56a, 56b | 57 | 58 | |
| Archival Storage - S3 (ID: 21a) | 59 | 60 | 61a, 61b | 62 | 63 | |
| Globus (ID: 21b) | 64 | 65 | 66a, 66b | 67 | 68 | |
| Research Computing Environment (ID: 10) | 69 | 70 | 71a, 71b | 72 | 73 | 74 |
| Cluster Filesystem (ID: 11) | | 75 | | 76 | 77 | |

**Figure 16. Portion of the STRIDE Table Demonstrating Format**

This figure facilitated threat tracking and can be used in conjunction with Table 12 to validate that all appropriate STRIDE threats were evaluated for each data element. The actual threats represented by the Threat IDs were in a separate table, with the first column being the Threat ID and the row describing the details of the threat.

### 2.2.2. Key STRIDE Results

The threats identified in the STRIDE analysis (see Table 4 and Table 5) were analyzed and prioritized based on real-world data and attacks that targeted the most valuable assets identified in Section 2.2. This section presents several examples of the "key" STRIDE threats while not implying that these threats are the only ones that need to be mitigated. All threats need to be addressed through elimination, mitigation, acceptance, or transfer. The process for addressing threats should consider likelihoods and impacts (ID.RA-05; MO:1,3,8) as well as any legal or regulatory requirements.

These ten key STRIDE threats were mapped to TTPs from the MITRE ATT&CK Framework for Industrial Control Systems (ICS) or Enterprise Systems, though this was not done for all threats identified. Threats that include MITRE ATT&CK TTPs have been observed being used by adversaries. Hypothetical threats and threats that have only been realized in a research setting are not included in the MITRE ATT&CK Framework.

In the following tables, each STRIDE threat is identified along with a brief explanation of how the threat could be exploited and why it could be particularly impactful. Table 4 describes the STRIDE threats specific to genomic sequencers and provides a link to the MITRE ATT&CK Technique for additional information.

**Table 4. STRIDE Threats Specific to Genomic Sequencers**

| STRIDE Threat and MITRE ATT&CK Technique | Description |
|---|---|
| **1. Genomic Sequencer Tampering via the Attached Workstation**<br>MITRE ATT&CK Technique(s):<br>T0884 – Connection Proxy | The Genomic Sequencer implicitly trusts the attached workstation. This trust relationship between the Genomic Sequencer and the attached workstation could be exploited to tamper with the Genomic Sequencer. |
| **2. Genomic Sequencer Tampering with Genomic Data on the Cluster Filesystem.**<br>MITRE ATT&CK Technique(s):<br>T0867 – Lateral Tool Transfer<br>T0884 – Connection Proxy | The Genomic Sequencer is generally trusted by the Cluster Filesystem, and this elevated privilege makes it a target for a malicious actor. If the Genomic Sequencer has read/write access it could tamper with (for example, encrypt during a ransomware attack) the entire datastore. An attack tree for this threat is shown in Figure 18. |

| STRIDE Threat and MITRE ATT&CK Technique | Description |
|---|---|
| **3. Local Secondary Storage Information Disclosure and Exfiltration**<br>MITRE ATT&CK Technique(s):<br>T0893 – Data from Local System | A malicious actor could obtain and exfiltrate data from the Genomic Sequencer's local secondary storage that temporarily stores sequence data during network outages |
| **4. Spoofing of Sequencer Management Control of the Lab Network Administration Interface**<br>MITRE ATT&CK Technique(s):<br>T0858 – Change Operating Mode | The communication between the Sequencer Management application from the Hosting Environment to the Genomic Sequencer's Lab Network Administration Interface allows for remote management of sequencer runs. For this threat, a malicious actor spoofs coming from the Sequencer Management application and sends malicious sequencer management instructions over a remote access connection to the Lab Network Administration Interface of the Genomic Sequencer. If these instructions are perceived as originating from the Sequencer Management application, they could disrupt the Genomic Sequencer's scheduled runs, disclose information about these runs, extract sequencing data, or even permanently incapacitate expensive laboratory equipment, including DNA sequencers |

Table 5 describes threats to the Genomics Sequencing Laboratory or Research Partner environments. Each threat has been mapped to MITRE ATT&CK Techniques for additional information.

**Table 5. STRIDE Threats to the Genomic Sequencing Laboratory or Research Partner Environments**

| STRIDE Threat and MITRE ATT&CK Technique | Description |
|---|---|
| **5. Tampering of Data from Research Partner Datastore by Bioinformatic Software**<br>**MITRE ATT&CK Technique(s):**<br>T1195 – Supply Chain Compromise | Bioinformatic software are vulnerable to malware injection. While vulnerability concerns exist in other domains, the bioinformatics tools supply chain is often funded through multi-year research grants, after which software maintenance is minimal, if completed at all, as developers move to other projects or career positions. Vulnerabilities such as poor sanitation of inputs and use of obsolete or insecure functions with known exploitations have been identified [14][15]. Bioinformatics code developers may also introduce vulnerabilities through insecure code re-use or by including dependencies that can be exploited. An attack tree for this threat is shown in Figure 18. |
| **6. Tamper or Exfiltration of Research Partner Data through Remote Access to a Datastore**<br>MITRE ATT&CK Technique(s):<br>T1078 – Valid Accounts<br>T1133 – External Remote Services<br>T1021 – Remote Services | Research environments need remote access to the internet to connect with a datastore (for example, Globus) and download software packages. Adversaries may connect to a Research Environment datastore through remote access. They may leverage valid accounts or open external remote services to tamper with and/or exfiltrate genomic data or subsequent analyses. |
| **7. Data Exfiltration from a Research Partner Datastore by Bioinformatic Software** | Genomic datastores, especially those that contain data from individuals with pharmaceutical-targetable diseases, are of significant value. Whole genome sequencing runs are costly (at least $1,000 per sample) and samples are difficult to obtain. The threat here is similar to what was previously described above as threat "5," though instead of ransomware, the motive for inserting |

| STRIDE Threat and MITRE ATT&CK Technique | Description |
|---|---|
| **MITRE ATT&CK Technique(s):** T1567 – Exfiltration over Web Service | malicious code into a bioinformatics software package could be data exfiltration. Common approaches to genome analysis often involve combining several bioinformatics software packages to process the data. Some software may also need external connections to reach external databases for additional information or resources. Thus, the software needs a connection to the internet, enabling an exfiltration threat. An attack tree for this threat is also shown in Figure 17. |
| **8. Exfiltrate or Tamper with Data in Transit from the Genomics Sequencing Laboratory Data Storage to Research Partner**<br>**MITRE ATT&CK Technique(s):** T1565.002 – Data Manipulation: Transmitted Data Manipulation | The data being sent from the Data Delivery DMZ to the Research Partner could be altered during transit, affecting data integrity. This could cause incorrect data to be used in downstream analysis and may even be done for commercial gain. A further risk is that data are not destroyed according to policy after being transferred by the sequencing provider. |
| **9. Spoofing User to Genomics Laboratory Data Storage**<br>**MITRE ATT&CK Technique(s):** T1078 – Valid Accounts | An actor could spoof that they are a trusted party connecting to Globus. Only authorized and authenticated users should be able to access the genomic data at the Genomics Sequencing Laboratory. However, adversaries have used methods to obtain valid credentials to spoof users. |
| **10. Spoofing Researcher to Research Partner Login**<br>**MITRE ATT&CK Technique(s):** T1199 – Trusted Relationship<br>T1078 – Valid Accounts | Research environments need proper authorization and authentication to protect patient consent, safeguard intellectual property, and prevent malware. Common IT behaviors in research environments, such as lack of MFA, sharing default passwords among research groups, multiple users sharing the same computer account, hardcoding passwords, or embedding credentials in software that is then shared into the public domain, are possible mechanisms that malicious actors could use to take advantage of trusted relationships. This TTP is possible through leveraging trusted relationships or obtaining valid credentials. |

### 2.2.3. Attack Trees

Although developing attack trees can require more expertise, they effectively tell the story of how threats can be exploited. They can help prioritize mitigations by helping those less skilled in cybersecurity understand the risks if mitigations are not implemented. The attack trees for this paper incorporated MITRE ATT&CK TTPs to provide details for some of the STRIDE threats because MITRE ATT&CK identifies specific threats that have been exploited by known adversaries. Attack trees help highlight effective mitigations and show how attacks often involve multiple steps and often leverage multiple threats. Attack trees can help identify when multiple TTPs are available to accomplish the next step in the attack tree, making the mitigation less valuable to the defender than mitigations that have only a single path to reach the next node.

Attack trees can help an organization understand the impact and likelihood of a cyber incident (ID.RA-05; MO:1,3,8) and prioritize threats by incorporating adversarial actions into threat modeling. Attacks typically require multiple steps, as documented in MITRE ATT&CK or the

Lockheed Martin Cyber Kill Chain®. This section outlines two attack tree examples, one for the Research Partner and one for the Genomic Sequencing Laboratory.

Attack trees incorporate various shapes and connections to provide context to a flow carried out by a malicious actor. The diagrams are organized in a top-down manner, meaning the top elements show possible starting points for the attacker to choose from. The bottom of the diagram identifies the ultimate goal, such as exfiltration of data or denial of service.

Squares represent TTPs from the MITRE ATT&CK Matrix that an attacker may leverage to reach their goal. Arrows leaving the squares indicate moving to the next step. Once a technique has been completed, the arrow either connects to the next technique or connects to a stadium, the "or" operator. The squares in the diagrams show areas where multiple TTPs are viable and where the flow can continue as long as at least one TTP is accomplished.

The diamonds contain conditional statements, evaluated as either true or false. If true, flow continues downward. If false, there may be additional TTPs that need to be carried out before making progress toward the end goal. The specific branch will be represented by "True" and "False" on the outward arrows from the diamond. Not all diamonds will have a branch for both possibilities. Some may only represent a true branch, signifying that the false condition results in no possible progress or alternatives.

#### 2.2.3.1. Attack Tree 1: Untrusted Software Implanted with Malware

Figure 17 illustrates how untrusted software can be used to conduct a ransomware attack and/or exfiltrate the genomic data of the Research Partner. Since the untrusted software's code is outside the analyst's control, preventing adversarial actions will need to take place after the code has been tainted. Possible mitigations for this attack include detecting that the code has been tainted or restricting its privileges by sandboxing or containerizing the code during execution. With these two mitigations in place, the code would have to be both stealthy to avoid detection and clever enough to detect containerization and escape. After this, the defender has several additional opportunities that can either prevent the malware from having the desired effect (such as firewalls that prevent ingress tool transfer or exfiltration of data) or detect the malware (for example, monitoring for elevation of privilege or encryption behavior) followed by a robust response to limit the damage. The listed TTPs of each step (Table 6) can be useful for evaluating whether a tool has coverage of that TTP using third-party coverage tests such as MITRE ATT&CK Evaluations.

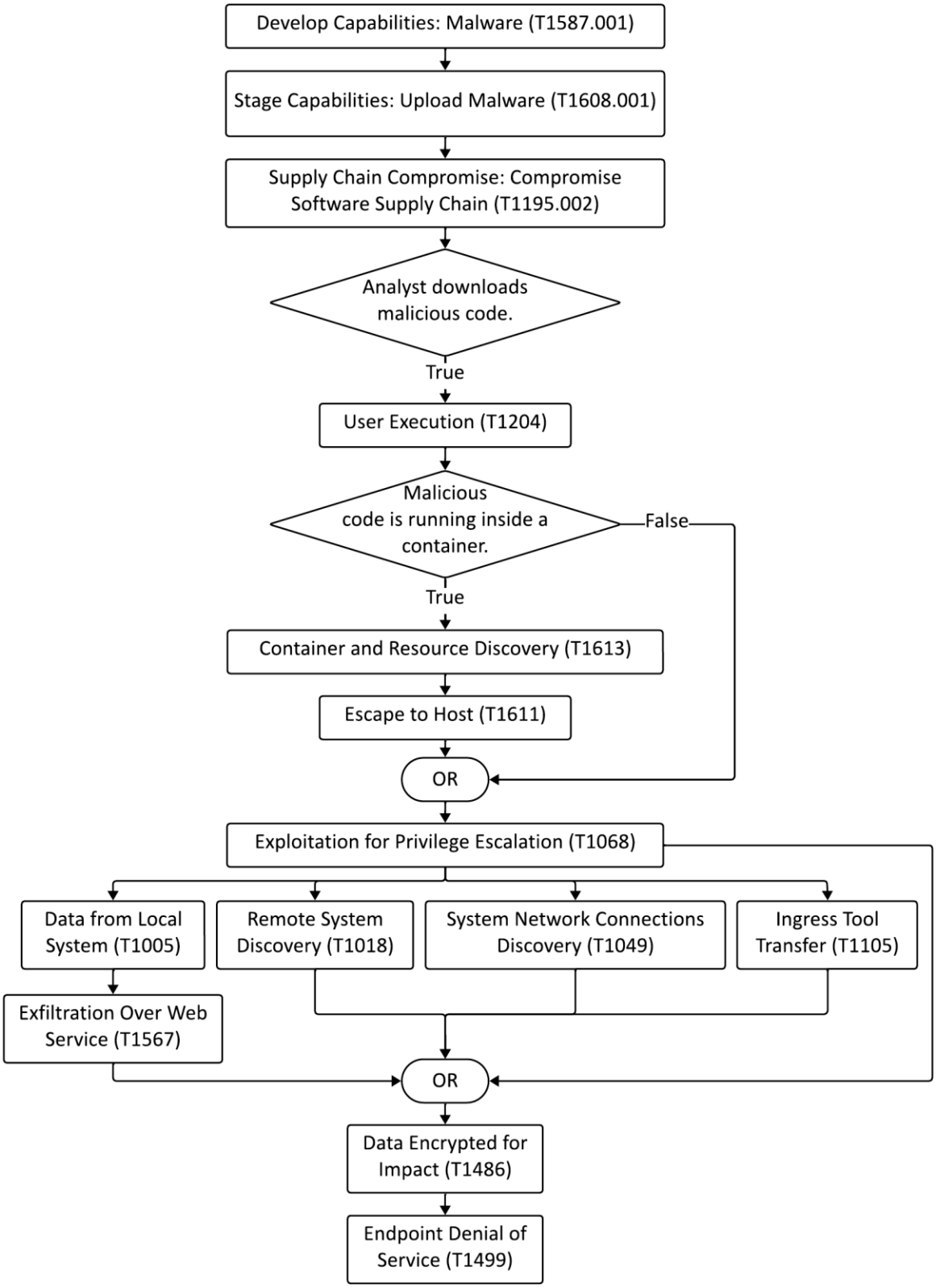



**Figure 17. Attack Tree 1: Untrusted Software Implanted with Malware**

**Table 6. Details for the Attack Tree 1**

| Technique ID and Name | Tactic ID and Name | Description |
|---|---|---|
| T1587.001 Develop Capabilities: Malware | TA0042 Resource Development | Attacker develops a new research tool containing malicious code. |
| T1608.001 Stage Capabilities: Upload Malware | TA0042 Resource Development | Attacker makes the research tool publicly available for analysts to download and use. |
| T1195.002 Supply Chain Compromise: Compromise Software Supply Chain | TA0001 Initial Access | Attacker forks a publicly available research tool and includes malicious code. |
| T1204 User Execution | TA0002 Execution | Analyst attempts to use the tool or model file they downloaded. |
| T1613 Container and Resource Discovery | TA0007 Discovery | Malicious code checks to determine whether it has been downloaded inside of a container. |
| T1611 Escape to Host | TA0004 Privilege Escalation | Malicious code utilizes incorrect container settings to escape and gain access to the host system. |
| T1068 Exploitation for Privilege Escalation | TA0004 Privilege Escalation | Malicious code gains additional privileges to attack the host system. |
| T1005 Data from Local System | TA0009 Collection | Attacker collects information from the compromised system. |
| T1567 Exfiltration Over Web Service | TA0010 Exfiltration | Attacker steals information to be used for blackmailing. |
| T1018 Remote System Discovery | TA0007 Discovery | Attacker scans for other systems to find options to move laterally. |
| T1049 System Network Connections Discovery | TA0007 Discovery | Attacker scans for network connections to find options to move laterally. |
| T1105 Ingress Tool Transfer | TA0011 Command and Control | Attacker installs a backdoor onto the compromised system. |
| T1486 Data Encrypted for Impact | TA0040 Impact | Attacker encrypts files on the host. |
| T1499 Endpoint Denial of Service | TA0040 Impact | Compromised system experiences denial of service event from the ransomware attack. |

#### 2.2.3.2. Attack Tree 2: Using the Genomic Sequencer Remote Access to Deploy Ransomware in Genomic Sequencing Laboratory Datastore

In this example, shown in Figure 18, the Genomic Sequencer is tampered with to gain access to the Cluster Filesystem in the Genomic Sequencing Laboratory environment. The manufacturer may have access to the sequencer to conduct updates and monitoring. Spoofing the manufacturer relationship would give the adversary significant access, not only to the device but also to other systems that trust the device.

This suggests that mitigations used to validate the manufacturer (such as enforcing firewall rules that only allow access from the manufacturer's servers, monitoring for brute force attacks, or restricting connections to Transport Layer Security (TLS) 1.3) could stop some of the attacks. Limiting the trust of the sequencer by the rest of the system could limit the damage of a sequencer compromise. If the Cluster Filesystem enforces partitioning of data from sequencers, such that each sequencer can only access its own data, and then only for a limited time, that could prevent ransomware from affecting the genomic data being stored on that system. Again, the listed TTPs of each step (Table 7) can be useful for evaluating whether a tool has coverage of that TTP using third-party coverage tests.

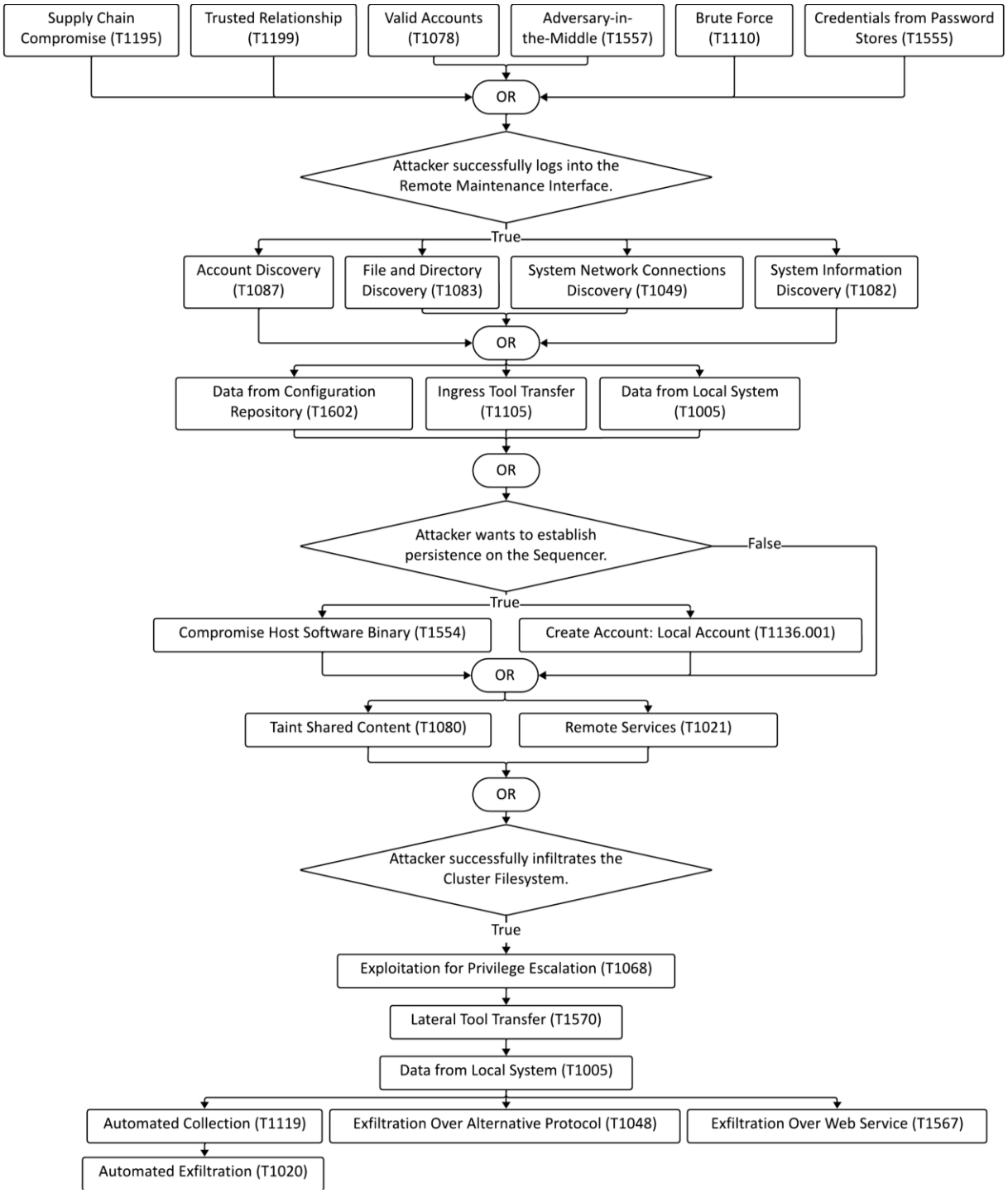



**Figure 18. Attack Tree 2: Using the Genomic Sequencer Remote Access to Deploy Ransomware in the Genomic Sequencing Laboratory Datastore**

**Table 7. Details for Attack Tree 2**

| Technique ID and Name | Tactic ID and Name | Description |
|---|---|---|
| T1195 Supply Chain Compromise | TA0001 Initial Access | System is compromised prior to being installed in the genomics laboratory. |
| T1199 Trusted Relationship | TA0001 Initial Access | Attacker compromises the device manufacturer and gains access. |
| T1078 Valid Accounts | TA0001 Initial Access | Attacker knows the credentials used by the manufacturer for remote maintenance. |
| T1557 Adversary-in-the-Middle | TA0006 Credential Access | Attacker intercepts network traffic to gain knowledge of the remote maintenance credentials. |
| T1110 Brute Force | TA0006 Credential Access | Attacker tries a brute force password attack to gain access to the device. |
| T1555 Credentials from Password Stores | TA0006 Credential Access | Attacker uses a well-known password to obtain access to the device. |
| T1087 Account Discovery | TA0007 Discovery | Attacker attempts to discover the accounts present on the Sequencer. |
| T1083 File and Directory Discovery | TA0007 Discovery | Attacker examines the directories and files present on the Sequencer. |
| T1049 System Network Connections Discovery | TA0007 Discovery | Attacker determines the network connectivity of the Sequencer. |
| T1082 System Information Discovery | TA0007 Discovery | Attacker tries to learn more about the operating system and services of the Sequencer. |
| T1602 Data from Configuration Repository | TA0009 Collection | Attacker collects information about how the Sequencer is configured. |
| T1105 Ingress Tool Transfer | TA0011 Command and Control | Attacker transfers a backdoor and additional tools onto the Sequencer. |
| T1005 Data from Local System | TA0009 Collection | Attacker collects information from the device/system. |
| T1554 Compromise Host Software Binary | TA0003 Persistence | Attacks creates a backdoor on the Sequencer. |
| T1136.001 Create Account: Local Account | TA0003 Persistence | Attacker creates a new local account on the Sequencer to allow for persistence. |
| T1080 Taint Shared Content | TA0008 Lateral Movement | Attacker alters run data and adds malicious code. |
| T1021 Remote Services | TA0008 Lateral Movement | Attacker exploits the connection between the Sequencer and Cluster Filesystem. |
| T1068 Exploitation for Privilege Escalation | TA0004 Privilege Escalation | Attacker exploits the Cluster Filesystem to gain escalated privileges. |
| T1570 Lateral Tool Transfer | TA0008 Lateral Movement | Attacker moves malicious tools from the Sequencer to the Cluster Filesystem. |
| T1119 Automated Collection | TA0009 Collection | Attacker sets up a mechanism to automatically collect run data from the Sequencer. |
| T1020 Automated Exfiltration | TA0010 Exfiltration | Attacker sets up a way to automatically exfiltrate the information collected. |
| T1048 Exfiltration Over Alternative Protocol | TA0010 Exfiltration | Attacker uses an available medium to exfiltrate the Cluster Filesystem data. |
| T1567 Exfiltration Over Web Service | TA0010 Exfiltration | Attacker uses a web service to exfiltrate the Cluster Filesystem data. |

## 2.3. Question 3: What are we going to do about it?

To address Question 3, the *Playbook* describes four strategies [9]:

1. **Eliminate.** This is the most desired outcome; however, it is often challenging and may involve forgoing a specific feature or functionality. For example, not collecting human subject health data would eliminate the threat of exfiltration of health data. If that feature or function is required to accomplish one of the scenario's Mission Objectives, then eliminating the threat is not possible.
2. **Mitigate.** This involves identifying, adding, and/or improving controls to protect, detect, respond to, or recover from attacks. For example, requiring multifactor authentication (MFA) instead of only username and password would mitigate (but not eliminate) the threat of someone spoofing an authorized user.
3. **Accept.** In any system, there are unmitigated threats that cannot be eliminated or mitigated whose risk is judged to be acceptable. However, these accepted threats need to be documented and periodically reviewed, as different organizations have different risk tolerance levels that may change over time.
4. **Transfer Responsibility.** This strategy transfers the risk to another entity, who may have resources of their own to mitigate the threat (for example, requiring users to choose secure passwords or documenting the risk in an informed consent agreement) or who are willing to accept the risk.

When working on Question 3, it is important to consider all four options: eliminate, mitigate, accept, and transfer. The impact on the mission posed by the threat, as well as the organization's risk tolerance, will guide decision-making. The most common and perhaps most complex option is to mitigate the threat using one or more mitigations to reduce the residual risk to an acceptable level. There may be multiple mitigations for a threat with varying costs and effectiveness. Choices of mitigations should be guided by the organization's mission, regulatory or legal requirements, risk tolerance, and resources.

Whichever mitigation options are chosen, they need to be documented adequately to be implementable. If a mitigation is called for, there should be sufficient detail so that the mitigation can be implemented and tested. Additionally, the remaining residual risk after the mitigation is implemented should be documented. If a risk is accepted, there needs to be sufficient documentation to understand the reasoning and assumptions that were used in deriving that solution because in the future, some of the assumptions may change, including an organization's risk tolerance.

This section explores mitigations that may address threats to genomic data, keeping in view the large data size, the need for secure sharing, and research environment requirements. The following list of mitigations, while not exhaustive, highlights key mitigations that emerged from the threat modeling exercise. A unique identifier for each mitigation was assigned to assist with documentation and traceability efforts. Identifiers for the Genomic Sequencing Laboratory (Lab) start with "L" while identifiers for the Research Partner (Partner) start with "P."

Table 8. Example Mitigation Table summarizes the mitigations detailed in the following sections. The table identifies the section number that describes the mitigation, the short title

for the mitigation and unique mitigation identifier (ID), the responsible party (Owner), the corresponding key threat number (from Section 2.2.2), the related attack tree (1 or 2), and the list of prioritized CSF Profile Subcategories along with applicable Mission Objectives (MO 1, 3, or 8 from Table 1).

In selecting and implementing mitigations, it is important to consider ownership, maintainability, verifiability (preferably automated), and usability. All systems will inevitably need updates and modifications. The responsible party (Owner) for maintenance and verification of each mitigation needs to be clearly defined. Mitigations should be verified after these changes to confirm that they still provide the expected utility. Systems can be monitored with appropriate logging and ongoing testing to identify any issues.

**Table 8. Example Mitigation Table**

| Section & Short Title (Unique ID) | Owner (Example) | ATT&CK Mitigation(s) ID and Name | Key Threat Number | Attack Tree | CSF Profile Subcategory and Mission Objective |
|---|---|---|---|---|---|
| 2.3.1 Broker Access (L1) | Lab IT | M1029 – Remote Data Storage<br>M1030 – Network Segmentation<br>M1035 – Limit Access to Resource Over Network | | | (PR.DS-01; MO:1,3,8)<br>(PR.DS-02; MO:1,3,8) |
| 2.3.2 Use Network Isolation and Firewalls (L2) | Lab IT | M1016 – Vulnerability Scanning<br>M1021 – Restrict Web-Based content<br>M1030 – Network Segmentation<br>M1037 – Filter Network Traffic<br>M1031 – Network Intrusion Prevention | | 1 | (ID.RA-02; MO:3)<br>(PR.DS-01; MO:8)<br>(PR.IR-01; MO:1,8) |
| 2.3.2 Use Network Isolation and Firewalls (P2) | Partner IT | M1016 – Vulnerability Scanning<br>M1021 – Restrict Web-Based content<br>M1030 – Network Segmentation<br>M1037 – Filter Network Traffic<br>M1031 – Network Intrusion Prevention | 6, 7 | 1 | (ID.RA-02; MO:3)<br>(PR.DS-01; MO:8)<br>(PR.IR-01; MO:1,8) |
| 2.3.2.1 Segment Network (L3) | Lab IT | M1030 – Network Segmentation | 4 | 1 | (PR.AA-05; MO:1) |
| 2.3.2.2 Firewall the Sequencer (L4) | Lab IT | M1035 – Limit Access to Resource Over Network<br>M1037 – Filter Network Traffic | 1, 3 | 1 | (PR.DS-01; MO:8)<br>(PR.DS-10; MO:1,8) |
| 2.3.2.3 Firewall the Cluster Filesystem (L5) | Lab IT and/or Cluster Filesystem Admin (shared) | M1035 – Limit Access to Resource Over Network<br>M1037 – Filter Network Traffic | | 1 | (PR.DS-01; MO:8)<br>(PR.DS-10; MO:1,8) |
| 2.3.2.4 Firewall the DMZ (L6) | Lab IT | M1035 – Limit Access to Resource Over Network<br>M1037 – Filter Network Traffic | | | (PR.DS-01; MO:8) |

| Section & Short Title (Unique ID) | Owner (Example) | ATT&CK Mitigation(s) ID and Name | Key Threat Number | Attack Tree | CSF Profile Subcategory and Mission Objective |
|---|---|---|---|---|---|
| 2.3.3 Use RBAC on the Cluster Filesystem (L7) | Cluster Filesystem Admin | M1018 – User Account Management<br>M1022 – Restrict File and Directory Permissions | 2 | 2 | (PR.DS-01; MO:8)<br>(PR.DS-10; MO:1,8) |
| 2.3.4 Authorize and Authenticate (L8) | Lab HR for Authorize; IT for Authenticate | M1018 – User Account Management<br>M1027 – Password Policies<br>M1032 – Multi-factor Authentication<br>M1036 – Account Use Policies | 1, 4, 9 | 2 | (GV.SC-02; MO:1,3,8)<br>(PR.AA-01; MO:1,3)<br>(PR.AA-03; MO:8)<br>(PR.AA-05; MO:1,3) |
| 2.3.4 Authorize and Authenticate (P3) | Partner Principal Investigator for Authorize; IT for Authenticate | M1018 – User Account Management<br>M1027 – Password Policies<br>M1032 – Multi-factor Authentication<br>M1036 – Account Use Policies | 6, 9, 10 | | (PR.AA-01; MO:1,3)<br>(PR.AA-03; MO:8)<br>(PR.AA-05; MO:1,3) |
| 2.3.5 Restrict Physical Access (L9) | Lab Security | N/A – ATT&CK does not cover physical mitigations | 1, 3, 4 | 2 | (PR.AA-06; MO:8) |
| 2.3.6 Implement Data Retention Policies (L10) | Lab Legal and Cluster Filesystem Admin | M1057 – Data Loss Prevention | 2 | 2 | (GV.OC-03; MO:1,3,8)<br>(ID.AM-08; MO:1,8) |
| 2.3.7 Conduct Backups (L11) | Admin for Cluster Filesystem; IT for Other Systems | M1053 – Data Backup | 2 | | (PR.DS-11; MO:1)<br>(PR.DS-01; MO:1,8) |
| 2.3.7 Conduct Backups (P4) | Bioinformaticist | M1053 – Data Backup | 5, 6 | 1 | (PR.DS-11; MO:1)<br>(PR.DS-01; MO:1,8) |
| 2.3.8 Containerize Untrusted Software (P5) | Bioinformaticist | M1048 – Application Isolation and Sandboxing | 5, 7 | 1 | (DE.CM-09; MO:1,3,8) |
| 2.3.9 Implement Least Functionality (L12) | Lab IT | M1033 – Limit Application Installation<br>M1042 – Disable or Remove Feature or Program<br>M1045 – Code signing<br>M1051 – Update Software | 2 | 2 | (ID.AM-08; MO:1,3,8)<br>(ID.RA-01; MO:3)<br>(PR.AA-05; MO:1,3,8)<br>(PR.PS-01; MO:1,3)<br>(PR.PS-02; MO:1,3,8) |
| 2.3.9 Implement Least Functionality (P6) | Bioinformaticist | M1033 – Limit Application Installation<br>M1042 – Disable or Remove Feature or Program<br>M1045 – Code signing<br>M1051 – Update Software | 5, 6 | 1 | (ID.AM-08; MO:1,3,8)<br>(ID.RA-01; MO:3)<br>(PR.AA-05; MO:1,3,8)<br>(PR.PS-01; MO:1,3)<br>(PR.PS-02, MO:1,3,8) |

| Section & Short Title (Unique ID) | Owner (Example) | ATT&CK Mitigation(s) ID and Name | Key Threat Number | Attack Tree | CSF Profile Subcategory and Mission Objective |
|---|---|---|---|---|---|
| 2.3.10 Encrypt Data (L13) | Lab IT and Cluster Filesystem Owner (Shared) | M1041 – Encrypt Sensitive Data | 2, 3 | 2 | (PR.DS-01; MO:1,3,8)<br>(PR.DS-02; MO:1,3,8) |
| 2.3.10 Encrypt Data (P7) | Bioinformaticist and Researcher (Shared) | M1041 – Encrypt Sensitive Data | 5, 6, 7, 8 | 1 | (PR.DS-01; MO:1,3,8)<br>(PR.DS-02; MO:1,3,8) |

### 2.3.1. Broker Access to Genomic Data

Organizations sharing genomic data may use an intermediary to provide protection between the internet and the Genomics Sequencing Laboratory datastore. A secure system that is hardened yet performant for transferring very large datasets is likely to meet these organizational needs. Setting up Globus in a Data Delivery DMZ can perform this intermediary function. A peer-reviewed design pattern for Globus setup is available at PeerJ Computer Science [16], though other implementations could also suffice. Key features the solution can provide include creating an intermediary between the untrusted internet and the storage system, enforcing strong authentication, encrypting data in transit, logging all access, and offering high performance.

### 2.3.2. Use Network Isolation and Firewalls

A target configuration for all firewalls used on the perimeter and within the environment is to deny all traffic by default and only allow the sources, targets, ports, and protocols required for functionality by the manufacturers or custom interconnectivity between networks. Specific protocols and destinations allowed include those needed for filesystem mounts, remote maintenance, vendor monitoring, software updates, and internal monitoring.

Details on how to properly secure the network are provided in NIST SP 800-215 [17]. Across all mitigations, configuration of firewalls, Domain Name System (DNS), and Network Time Protocol (NTP) should follow these recommended practices:

- Firewalls should use external dynamic lists (EDLs) that block known malicious sites before allowing exceptions.[7] These EDLs need to be configured to update automatically (e.g., daily) since threats are constantly evolving. An organization may also choose to block all access from locations of concern.
- DNS should conform to guidance on Protective DNS if available, as recommended by the National Security Agency (NSA) and the Cybersecurity & Infrastructure Security Agency

[7] Example EDLs are available at https://docs.paloaltonetworks.com/pan-os/10-1/pan-os-admin/policy/use-an-external-dynamic-list-in-policy/built-in-edls and https://rules.emergingthreats.net/.

(CISA) [18]. If that option is not available or considered too costly to implement, DNS can use free alternatives such as Quad9, OpenDNS, or Google Public DNS that provide blocking for known malicious domains.

- NTP should only allow for an internal time server and use only approved time services, such as NIST's Official U.S. time service or similarly trusted time service.

The subsections below describe mitigations, including network segmentation and specific exceptions needed for the firewalls that protect the Genomic Sequencer and the Cluster Filesystem. Some of the devices, such as the Genomic Sequencer, Cluster Filesystem, and servers, may include firewalls. These firewalls can provide defense in depth and can be activated and configured to limit traffic, though this is not a substitute for more sophisticated firewalls with EDLs along with advanced capabilities at the perimeter. Advanced firewall capabilities can provide perimeter protection, threat prevention, and network analysis of lab instrumentation, Internet of Things (IoT) devices, and security enclave infrastructure.

The applications and services within the Management and Tooling environment could include several passive and active security tools to protect the research and laboratory environments. Vulnerability scanning can be performed on a scheduled and *ad hoc* basis to detect security flaws in underlying hardware and software within the security enclave. Behavioral analysis and threat detection could be performed on the logged ingress and egress network traffic within the secure enclave. Network intrusion detection sensors could provide real-time alerting of suspicious activity within the enclave. In addition, all network traffic in the enclave could be recorded and retained for a designated period to enable in-depth analysis of device and user activities.

#### 2.3.2.1. Network Segmentation

Different sections of the Genomic Sequencing Laboratory can be segmented from each other using virtual local area networks (VLANs). At minimum, the Wet Lab, the Research Environment, the Data Delivery DMZ, and the Management and Tooling environments are expected to be separate zones. These zones are ideally segregated on different networks or virtual networks with explicit, limited access between the zones using access control lists (ACLs). Data traveling between zones can use the most recent version of TLS (e.g., TLS 1.3).

Trust boundaries like the Research Computing Environment, hosting environment, and Data Delivery DMZ are VLANs that are segregated from other environments and the outside world. Laboratories and sequencing networks have explicit permissions that allow them to connect and execute genomic pipelines in the security enclave. This is accomplished by segmenting VLANs, implementing ACLs, and using network isolation tools and firewall practices mentioned in the previous section.

#### 2.3.2.2. Firewalls for the Genomic Sequencer

Firewalls on and around the Genomic Sequencer (such as its attached workstation) can be configured to deny-all traffic by default, allowing only the ports and protocols required by the

manufacturer to protect from threats like those described in Section 2.2.2 and attack tree 2 (Figure 18). Specific protocols and destinations that can be allowed include access to the Cluster Filesystem mounts, remote maintenance, vendor monitoring, software updates, and internal monitoring. Organizations can follow manufacturer guidance on allowed network connections. For example, Table 9 lists the endpoints for ingress and/or egress to support the Illumina sequencer[8], sorted by geographic region.

**Table 9. Illumina ACLs**

US East (N. Virginia)

2. Identify the endpoints for your instrument.
Each instrument includes specific endpoints that are categorized as either required, recommended, or optional. These endpoints are used for the following purposes:
- Authorizing certificates
- Displaying fonts
- Telemetry
- Accessing Illumina support material
- Sending IDAT files or data to ICA

iScan

The following table shows the applicable endpoints for the iScan.

| Endpoint | Category | Purpose |
|---|---|---|
| ica.illumina.com | Required | Send IDAT files to ICA |
| o.ss2.us | Required | Certificate authorization |
| ocsp.digicert.com | Required | Certificate authorization |
| ocsp.pki.goog/gsr2 | Required | Certificate authorization |
| ocsp.rootca1.amazontrust.com | Required | Certificate authorization |
| ocsp.rootg2.amazontrust.com | Required | Certificate authorization |
| ocsp.sca1b.amazontrust.com | Required | Certificate authorization |
| fonts.gstatic.com | Required | Display fonts |
| fonts.googleapis.com | Recommended | Display fonts |
| cdn.walkme.com | Recommended | Telemetry |
| cdn3.userzoom.com | Recommended | Telemetry |
| dpm.demdex.net | Recommended | Telemetry |
| illuminainc.demdex.net | Recommended | Telemetry |
| illuminainc.tt.omtrdc.net | Recommended | Telemetry |
| smetrics.illumina.com | Recommended | Telemetry |
| google.com | Recommended | Telemetry |
| google-analytics.com | Recommended | Telemetry |
| stats.g.doubleclick.net | Recommended | Telemetry |
| illumina.com | Optional | Access Illumina support material |

Table 10 provides another example of the PacBio Sequencer ACLs [19][20]. Additional information is available at their support site, including preparation documentation for ports and firewalls.

[8] More information available at: https://support-docs.illumina.com/SHARE/NetworkSecurity/Content/SHARE/NetworkSecurity/ControlComputerFirewall.htm.

**Table 10. PacBio Sequencer ACLs**

| Source | Destination | Port/Protocol | Description |
|---|---|---|---|
| Revio Instrument Control Computer (ICC) | SecureLink Servers | 443/tcp | Communication for remote support (PacBio Insight) |
| ICC | Data Transfer Server | 22/tcp, 873/ tcp, or 80/tcp and 443/tcp depending on protocol | Data transfer from instrument to customer storage |
| ICC | Customer or external NTP servers | 123/udp | Used for updating machine time. Defaults to pool.ntp.org |
| ICC | Customer server | 53/udp or 53/tcp | Nameservers |
| ICC | SMRT Link server | 8243/tcp | Communication from instrument to SMRT Link |
| SMRT Link server | ICC | 9243/tcp | Communication from SMRT Link to instrument |
| Customer laptop/ desktop PC | SMRT Link server | 9090/tcp | SMRT Link GUI http |
| Customer laptop/ desktop PC and ICC | SMRT Link server | 8243/tcp | SMRT Link web services and GUI https |
| Customer laptop/ desktop PC | SMRT Link server | 9443/tcp | SMRT Link Administration https (API Management Interface) |
| SMRT Link server | Shared Network File System (NFS) storage | NFS ports (may vary depending on configuration) | NFS shared storage access shared data to analyze |
| SMRT Link server | PacBio Event server (https://smrtlink-eve.pacbcloud.com:8083) | 8083/tcp | Optional reporting of server metrics to PacBio Tech Support |
| SMRT Link server | PacBio Update server (https://smrtlink-update.pacbcloud.com:8084) | 8084/tcp | Downloading Chemistry Updates |
| HPC nodes | Shared NFS storage | NFS ports 2049/ tcp (ports vary depending on configuration) | NFS shared storage access shared data to analyze |

Generally, the sequencer will communicate with a Cluster Filesystem or other storage using industry-standard mount protocols like Network File System (NFS) and Server Message Block (SMB) along with varying types of storage from block, object, or database access, all with their necessary ports and protocols. Additional ports and protocols may be required to allow the sequencer to access the internet for updates and vendor support. These ports and protocols need to be allowed, with proper restrictions in place for specific source and destination addresses.

#### 2.3.2.3. Firewalls for the Cluster Filesystem

Cluster Filesystem firewalls can also be configured with a default deny-all rule, allowing only the ports and protocols required by the Cluster Filesystem (e.g., Globus, sequencers, researchers in the zone). Generally, a Cluster Filesystem will also support industry-standard mount protocols like NFS and SMB along with varying types of storage from block, object, or database access, all with their necessary ports and protocols. Refer to manufacturer guidance for more details.

#### 2.3.2.4. Firewalls for the Data Delivery DMZ

Data Delivery DMZ firewalls can be configured appropriately to mitigate the significant risks from outside connections. Advanced firewall capabilities can be leveraged to protect perimeter networks and internal network traffic connecting trust boundaries. Capabilities that can be enabled include advanced firewall functions such as packet filters and network address translation, stateful inspection, deep packet inspection, threat prevention, audit and logging, and access control.

The Data Delivery DMZ will require access to ports 80, 443, 4443, 50000, and 51000 inbound from ANY and from those same ports outbound to ANY (note, Section 2.3.2 EDL rules need to be executed before this exception, will deny connections from known malicious IP addresses, and can be configured to also block IPs from locations of concern). Ports 50000 to 51000 are used for GridFTP data channel traffic and used only during transfers as needed; the data channel traffic is sent directly between endpoints and not the Globus service. Port 443 inbound is used by the manager service, GridFTP control channel traffic, and Hypertext Transfer Protocol Secure (HTTPS) access to collections. Port 443 outbound is used to communicate with cloud storage services, pull Globus Connect Server packages from the Globus repository, and communicate with the Globus service through its representational state transfer (REST) application programming interface (API).

### 2.3.3. Use RBAC on the Cluster Filesystem

Organizations will benefit from the maintenance and enforcement of ACLs across all internal networks, leveraging role-based access control (RBAC). This includes all connections requiring X.509 certificates to be permitted for internal IP addresses only. Access to the management APIs or web console should be validated against either an external Lightweight Directory Access Protocol (LDAP), Active Directory, or OpenStack Keystone data and allowed only with MFA. The Cluster Filesystem in the Research Computing environment can encrypt data at rest and in transit. Figure 19 shows an example of client certificates that can also be used to limit access for specific roles to only allow read-only or limit privileged access. RBAC can limit sequencer access to a specific directory, separate from other sequencers and data that have been quality controlled.

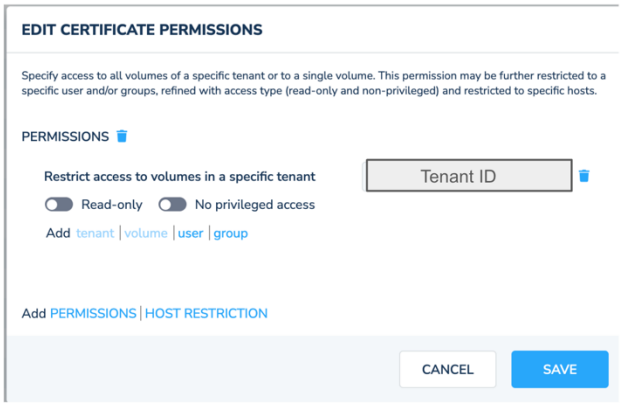



**Figure 19. Certificate Example**

### 2.3.4. Authenticate and Authorize All Users

Enforcing authentication and authorization will reduce risk to all environments for both the Research Partner and Genomic Sequencing Laboratory. This can be done by requiring unique users, promptly revoking credentials of users that depart an organization, RBAC, strong passwords, lockout on too many failed login attempts, and MFA. Local and remote user access can be controlled using a combination of LDAP, Kerberos, Single Sign-On (SSO), and network ACLs. These systems limit and control access to the security enclave, specific systems, and volumes based on the user's role. Remote users, including genomic and cybersecurity providers, can connect to the security enclave using their unique credentials, MFA, and a virtual private network (VPN such as Global Protect) client. Organizations can implement MFA with an authenticator service such as Google Authenticator, Duo Authenticator, 2FAS[9], or RSA Authenticator.

There will be cases where, for some users, the authentication and authorization responsibilities are transferred. For example, the Genomic Sequencing Laboratory may require the Research Partner to enforce the authentication and authorization requirements for the users that will have access to the Research Partner's data in the Cluster Filesystem. The transfer of that responsibility needs to be communicated from the Genomic Sequencing Laboratory to the Research Partner.

### 2.3.5. Restrict Physical Access to Environments

Facilities hosting research computing and hosting environments can benefit from performing a risk assessment to determine the layers of physical security needed to protect personnel, assets, and data. The risk assessment may lead to recommending the following mitigations:

- Security personnel to check identity badges along with access-control doors for rooms where servers are located
- All points of ingress and egress to be controlled using automatically locking doors equipped with key card readers, supplemented by a 24/7 security guard presence and surveillance cameras
- Access to more sensitive areas like the data centers and cabinets to be controlled via key cards, multi-factor passcodes, and mantraps to restrict access to only those with a legitimate business need
- Additional video surveillance, fencing, and physical security protections may also be needed around the facilities
- Prompt revocation of physical access for users who leave the organization

### 2.3.6. Implement Data Retention Policies for the Genomic Data

Data retention policies for organizations processing genomic data will vary based on contractual and regulatory requirements. Contractual requirements should be straightforward, written into

[9] 2FAS is an open-source two-factor authentication (2FA) tool.

formal agreements, and regularly updated. Regulatory requirements can be more complicated. The source of the data may impact the retention requirements. For example, human genomic data will fall under different regulations than non-human genomic data, such as privacy and consent requirements. The intended use of the data may also impact retention requirements. Data collected for research purposes will have different requirements than data collected in a clinical context for use in treatment or diagnosis. Once the retention period has elapsed, genomic data can be deleted or moved to offline storage to limit liability and prevent exfiltration, depending on contractual and regulatory requirements. Care should be taken with any remaining physical samples, which may need to be destroyed or returned to the Research Partner.

### 2.3.7. Conduct Backups of Datastores

A variety of backup options are available for sequencing data at both the Research Partner and the Genomic Sequencing Laboratory. These range from keeping DNA in the freezer to re-sequence when needed to back up sequencing files in a variety of genomic data formats. Examples of backup formats include the file format for sequence reads with quality score (FASTQ), Binary Alignment/Map (BAM), and Compressed Reference-oriented Alignment Map (CRAM).

Table 11 details examples of backup options, provides encryption times, and estimates storage costs per year for an offsite cloud backup as calculated at the time of this document (2024). For Table 11, file sizes are in gigabytes (GB), times are in minutes (m) and seconds (s), and cloud storage uses Amazon Web Services (AWS) S3 Deep Glacier Flexible Retrieval.

An important aspect of backing up genomic sequencing data is the ability to recover to the last known good state, as each file storage option will restore to a different analysis state. A genomic analyst will need to identify where in the genomic data lifecycle their analysis or responsibilities exist to determine the appropriate backup option. Periodic validation for functional backup checks could include re-mapping or variant calling retrieved backup data, depending on an organization's need upon recovery of backups.

**Table 11. Genomic Sequencing Backup Options**

| Backup Option | File Size (GB) for 30x Human Genome on Illumina NovaSeq | Encryption Time | Notes on "Last Known Good State" | Cost per Year for Secondary Backup in AWS S3 Deep Glacier Flexible Retrieval |
|---|---|---|---|---|
| DNA in Freezer – sequence as needed | N/A | N/A | Pros – researcher might get to sequence on new technology<br>Cons – cost is likely higher than data backup, not applicable for limited material | Freezer maintenance costs likely amortized over samples and other research projects |

| Backup Option | File Size (GB) for 30x Human Genome on Illumina NovaSeq | Encryption Time | Notes on "Last Known Good State" | Cost per Year for Secondary Backup in AWS S3 Deep Glacier Flexible Retrieval |
|---|---|---|---|---|
| FASTQ.GZ (compressed reads file) | 2 Files<br>Reads 1 – 24GB<br>Reads 2 – 25GB | Reads 1 –<br>real 14m15.653s<br>Reads 2 –<br>real 16m15.559s | Requires re-mapping of reads that may be expensive if needed to perform for many samples | $2.12 |
| BAM (mapped reads file) | 37GB | real 22m30.975s<br>user 1m11.776s<br>sys 1m27.350s | May be the easiest to work from, but it locks a user into a reference genome and could require extra work | $1.60 |
| CRAM (compressed mapped reads file) | 14GB | real 8m8.965s<br>user 0m25.891s<br>sys 0m28.499s | Not as many analysis tools use a CRAM file as input compared to a BAM, so the user will need to know the impact on their pipeline | $0.60 |

### 2.3.8. Containerize Untrusted Software

A mitigation to the threat from untrusted bioinformatics software is to run this software in containers with restricted privileges and access. Damage from implanted malware surreptitiously included in open-source analysis packages can be limited by using hosts and containers employing a mandatory access control system, whereby process access is controlled by the system. For example, AppArmor can be used within containers to limit the access of running processes, restricting what files they are allowed to access and what types of actions they may perform on these files. Judicious configuration of AppArmor profiles can restrict which files a process may execute, mitigating or eliminating the impact of implanted malware.

### 2.3.9. Implement Least Functionality and use Configuration Benchmarks

Benchmarks and least functionality are best practices that can be enforced across all environments. Least functionality will help eliminate potential risks resulting from running unneeded services that may be leveraged by adversaries. Organizations can use configuration benchmarks such as the Defense Information Systems Agency (DISA) Security Technical Implementation Guides (STIGs) or Center for Internet Security (CIS) Benchmarks when available for each component of the network. If manufacturers provide configuration benchmarks or guidance, these can also be enforced. This mitigates the threats from unneeded remote services running on the system, as described in STRIDE Threat 6 in Section 2.2.2. To maintain software, teams can enable automatic security updates and scan regularly, remediating discovered vulnerabilities as described in the *Genomic Data Profile* Subcategories (ID.RA-01; MO:3).

### 2.3.10. Encrypt Data Whenever Possible

One of the most effective ways to protect data is encryption. The CSF includes Subcategories for protecting data at rest and in transit (PR.DS-01, PR.DS-02; MO:1,8).

Encryption at rest is accomplished via command line by executing "`$openssl enc AES128-CBC`" on the server before sending the data to secondary storage in a cloud bucket [21]. Wall clock encryption times using OpenSSL with these parameters range from 16 minutes and 15 seconds for a 25GB FASTQ.GZ reads file to 8 minutes and 9 seconds for a 14GB CRAM compressed alignment file.

Encryption of data-in-transit between the Genomics Sequencing Laboratory and the Research Partner is accomplished using the Globus "encrypt transfer" option.

When transferring data between two endpoints using Globus, a "data channel" is established directly between the source and destination endpoints. The data channel is inaccessible to the Globus service but can be accessed by the servers running the endpoints. Users initiating the transfer can choose to encrypt the data channel, or the endpoint administrator can enforce the encryption of all transfers to or from an endpoint. The specific cipher used for a transfer is negotiated between the source and destination endpoints based on their preference-ordered list of OpenSSL ciphers (default HIGH). Additionally, Globus employs a TLS-encrypted "control channel" to communicate with the source and destination endpoints during transfers.

Many Cluster Filesystems will support encryption in transit to endpoints using client certification permissions like those shown in Figure 20 [22], encrypting and limiting access to the Filesystem to end users and specific systems.

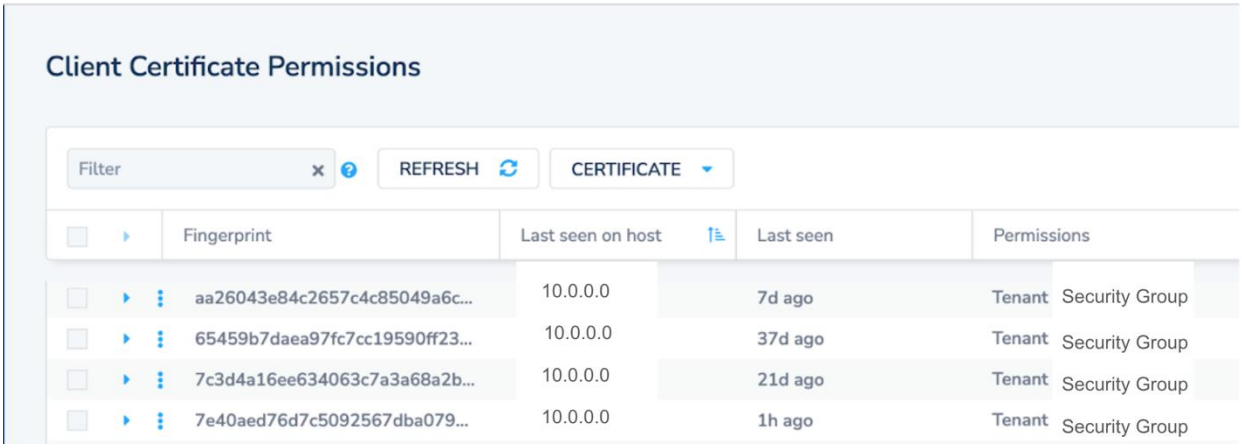



**Figure 20. Client Certificate Permission List of a Cluster Filesystem**

## 2.4. Question 4: Did we do a good job?

Question 4, "Did we do a good job?" directs the team to evaluate the effectiveness of answers to Questions 1–3. This paper outlines the efforts to document the genomic data processing environments (Question 1), identify threats (Question 2), and implement mitigations (Question 3). The threat modeling process is designed to be iterative. This paper does not attempt to be

comprehensive, but rather to demonstrate the process so that other teams can leverage this work to conduct their own threat modeling. Question 4 helps emphasize that this process will be repeated to address changes in the system and threat environments.

This section is designed to describe a concrete example of how to address the question, "Did we do a good job?" and provide additional activities that can be used by teams to evaluate their efforts. All documentation should be easy to update and reviewed periodically to address new vulnerabilities, system changes, new assumptions, and changes in risk tolerance.

### 2.4.1. Did we do a good job documenting the system and data architecture?

Section 2.1 documents DFDs and HVDs that deserve special attention in consideration of threats against and mitigations of the threats due to their nature of crossing trust boundaries and/or affecting critical systems. To provide examples for this process, several of the identified HVDs were modeled in more detail with cross-functional diagrams.

The following activities could be used to improve the documentation of the system and data architecture:

- Analyze HVDs that have not been documented.
- Review documentation and information from suppliers, developers, and users to consider any updates required.
- Review change control processes to ensure that new devices are captured and other changes are documented properly.
- Review personnel onboarding and offboarding processes.
- Review network segmentation and firewall configurations to ensure compliance with best practices.
- Update the documentation to reflect changes to the threat or system environment, including system interconnections, devices added, configurations, access controls, or issues identified through testing or monitoring.

The following additional questions help evaluate "Did we do a good job?" answering Question 1, "What are we working on?"

- Is the DFD sufficiently detailed to capture communications between systems, particularly those that cross trust boundaries?
- Are all communications that cross trust boundaries included?
- Have HVDs been highlighted and is there sufficient model detail (such as cross-functional, state, and swim lane diagrams) to understand threats against them?
- Does the threat modeling explain how HVDs work and assess the impact of threats and mitigations on them? Are the diagram details sufficient, or is additional information needed from suppliers, developers, or users?

- Are the DFD's trust boundaries accurate? Can they be enforced (for example, by network segmentation)?
- Is there a justification for every "allow" network firewall rule from a device? For Question 1, this is not to evaluate the mitigation but to be sure that you have mapped all the dataflows.
- Is there a trust boundary for every control mechanism (such as firewall, ACL, lock, or login)? Where there is a control mechanism, it likely represents a trust boundary.

### 2.4.2. Did we do a good job identifying and documenting threats?

To answer, "Did we do a good job?" on Question 2, "What could go wrong?" the project team evaluates whether the threat model adequately identifies and documents threats to the system. Section 2.2 enumerates the threats identified from the STRIDE analysis and the attack trees. The team created a table of threats with unique identifiers and selected "key" threats for further analysis and specific mitigations. The team highlighted high-value resources such as the genomic datastores and the sequencer to focus initial threat identification efforts. The team reviewed the STRIDE Element chart and the attack trees to consider gaps in the initial threat identification process.

The following actions could improve threat identification:

- Review organizational policies, strategies, and processes to determine if there are other threat areas not being addressed by the technical evaluation.
- Address additional missing STRIDE Elements based on Table 12.
- Develop additional attack trees to address broader threat scenarios.
- Review published threats and actual cyber incidents identified that are targeted toward genomic data to verify they are included in the project's threat table or attack trees.
- Incorporate privacy threat modeling to address potential genomic data privacy concerns.

The following additional actions help evaluate "Did we do a good job?" answering Question 2, "What could go wrong?"

**Evaluate the comprehensiveness of the STRIDE analysis.** The STRIDE methodology has an effective completeness check that uses the STRIDE per element mapping shown in Table 12.[10] With this table, a completeness check can be done for the typical threats against external entities, processes, datastores, and dataflows. The "X" in Table 12 indicates what threats should be present, while the absence of an "X" indicates threats that are not considered and a "?" indicates that it depends on the details whether it could be present. For example, in STRIDE, Tamper threats against an external entity are not considered because they are outside the scope of the organization's knowledge and control (see the *Playbook* [9] for details).

[10] Reproduced from the *Playbook* [9].

**Table 12. STRIDE per Element**

| *Element* | *Spoof* | *Tamper* | *Repudiate* | *Info Disclosure* | *DoS* | *EoP* |
|---|---|---|---|---|---|---|
| *External Entity* | X | | X | | | |
| *Process* | X | X | X | X | X | X |
| *Data Store* | | X | ? | X | X | |
| *Dataflow* | | X | | X | X | |

If there are expected threats that have not been considered by the team, the absence of information on that element in the system's threat table highlights an area for additional consideration. For example, if a repudiation threat against an external entity is not identified in the threat table, that would indicate a gap. Threats against HVDs should be revisited to determine if additional review is necessary or if applying more than one method (e.g., STRIDE and an attack tree) may be helpful.

**Evaluate the comprehensiveness of the attack trees.** When evaluating attack tree documentation:

- Consider attacks that have occurred in the genomic stakeholder community and closely adjacent industries. Threat intelligence can be used to identify TTPs favored by actors who are known to target an industry.
- Consider whether the attack trees reflect these attacks, or if additional attack trees should be developed.
- Consider known vulnerabilities in software and services being used.
- Determine whether the threats being considered map to the threats listed in published documents for the genomic community, such as NIST IR 8432 [3].
- If the system is operational, consider if past downtime can be mapped to threats identified in the threat model.

### 2.4.3. Did we do a good job mitigating the threats?

Section 2.3 documents the mitigations considered as part of this threat model. Specifically, the team focused on ten "key" mitigations that addressed numerous threats identified, tailoring them to the genomics data sequencing workflow use case. Table 8 maps these key mitigations to the key STRIDE threats identified, the two attack trees, and CSF Subcategories prioritized from the *Genomic Data Profile*.

The following actions could evaluate and improve on these initial mitigations:

- Review the mitigations to assess how well they address the key threats identified from STRIDE and the two attack trees.
- Expand mitigations to cover additional CSF Subcategories (such as Govern, Respond, and Recover) from the *Genomic Data Profile* that may not be captured in the initial threat model analysis.

- Review the documentation from Question 1 to ensure that all mitigations are included.
- Review any changes made to Question 1 and Question 2 to identify additional potential mitigations needed.
- Expand the mitigations to cover additional threats identified (including those from additional attack trees created) and controls prioritized by the CSF Profile Subcategories.
- Develop a mitigation monitoring plan that incorporates any findings from assessments, tabletop exercises, or ongoing monitoring and documents how they will be integrated into future threat modeling activities.

The following additional actions help evaluate "Did we do a good job?" answering Question 3, "What are we going to do about it?" These activities help evaluate the thoroughness of mitigations and regularly consider the impact of any changes to the system or threat environment. A legal review may be appropriate to determine if the mitigations, accepted risks, and transferred risks (particularly the manner of transfer notification) meet the necessary regulatory requirements (GV.OC-03; MO:1,3,8).

**Review Risk Strategies.** Determine if there is a risk strategy for every threat that crosses a trust boundary and consider mitigations and other responses across each risk strategy, such as eliminate, accept, and transfer.

- **Eliminate.** Eliminating threats often removes features. Whenever threats are eliminated, documentation should justify why the risk from a threat was deemed to outweigh the benefit from the feature. This documentation is necessary because threat models will need to be revisited as the system and organization evolves. Future threat modeling efforts may involve different participants who may not be familiar with the system and will rely on documentation.
- **Accept.** Threats that are risk-accepted should be documented sufficiently to explain why the risk was accepted. For example, an authentication threat may be accepted because there is no remote login and there are physical controls restricting access to a device. The reason for the risk acceptance needs to be documented so that if the system is modified to allow remote access or moved to a place without physical controls, it will be clear that the risk needs to be reassessed.
- **Transfer.** When threats are transferred, complete and sufficient documentation fully assigns responsibility to the entity assuming accountability for the risks that derive from that threat. That entity may then also be responsible for accepting, mitigating, or transferring the risk. For example, responsibility for authorizing and authenticating Research Partner users who can access the Genomic Sequencing Laboratory data could be transferred to the Research Partner. Does the documentation adequately inform the Research Partner of their responsibility and define what the required authentication mechanism is (such as username and password or MFA)?

**Update DFDs.** As mitigations are added, DFDs may need to be updated. Threats against that element should be considered, and a risk strategy should be assigned to eliminate, mitigate, accept, or transfer. For example, if you add a firewall, the DFD should be updated to include the

firewall (a firewall may have an administration console, a configuration file, etc.) and then evaluate the threats against the firewall.

**Review Attack Trees.** Attack trees are a helpful tool in addressing the question, "Did we do a good job?" particularly when they are based on methods or TTPs that have been known to be in use by adversaries. If there are mitigations in place that sever the attack tree in multiple places, that can be a positive indication of the layering of controls which can be part of a robust cybersecurity defense.

**Use CSF Profiles.** Teams can use the *Genomic Data Profile* to identify additional mitigations by considering priority Subcategories for each Mission Objective. The mitigations selected during Question 3 activities can be mapped to CSF Subcategories and used to develop a CSF Profile tailored to the organization. The organization can identify potential gaps by comparing the organization's CSF Profile to an appropriate target profile like those provided by the *Genomic Data Profile* [4].

**Track Mitigations Throughout the System Lifecycle.** Threat mitigations should be documented, reviewed, tested, and maintained as the threat environment changes. This may include the following considerations:

- During the implementation phase, the threat modeling should be periodically revisited and updated. Consider whether the mitigations caused problems and if so, what were the impacts.
- Once mitigations are operational, consider their effectiveness and any negative impact to Mission Objectives. During security incident response and recovery, determine if mitigations increase or decrease system uptime (ID.IM-03; MO:1,3,8). If the mitigation decreased system uptime, consider if the protection provided by the mitigation justified the loss of system availability.
- Organizations should update their threat model after a device or mitigation fails. While device failures are often unrelated to cybersecurity issues, they can be useful for evaluating resiliency measures, which are an important part of response and recovery from cyber incidents.
- Organizations should update their threat model when any significant modifications are made to the system.
- Security assessment, including automated and manual penetration testing, is another useful tool to evaluate how the mitigations and threat modeling perform and how they can be improved.
- Tabletop and Functional Exercises as described in NIST Special Publication (SP) 800-84 [23] can also be very helpful in evaluating Question 3 of the threat modeling process and can be done both before and after a system is in use (ID.IM-02; MO:1,3,8).

## 3. Conclusion

This paper demonstrates cybersecurity threat modeling techniques to evaluate potential threats for the common genomic data processing use case where:

- An organization (Research Partner) sends a physical DNA sample and associated metadata (in digital form) to a genomic sequencing provider (Genomic Sequencing Laboratory).
- The Genomic Sequencing Laboratory generates genomic data from the physical sample, processes the data, and sends the results to the Research Partner.
- The Research Partner then analyses the data using tools that include untrusted software and publicly available reference data.

This paper provides an example of how a threat modeling process can be employed in a systematic and consistent manner to analyze threats to the Research Partner and Genomic Sequencing Laboratory environments. It shows how the process identifies dataflows in each environment and highlights the high-value dataflows that may warrant additional mitigations. It also identifies and characterizes some key threats against these environments and describes sample mitigations genomic data processing organizations may consider.

This threat modeling process identified three areas where genomic data processing concerns and threats differed from most enterprise applications:

- Protections to address the unique value and potential size of genomic data, including closely managing remote access when sharing with external partners
- Controls to protect the data when running untrusted researcher code during genomic data analysis
- Safeguards to protect the highly valuable sequencers in the internal network that process sensitive genomic data and provide manufacturers access for maintenance

Organizations that process genomic data can use this paper to guide them in conducting threat modeling on their own unique environments. The threat model results can be used to:

- Guide system development and implementation choices to mitigate threats to the organization.
- Document how a system is intended to function, threats to the system, and strategies to address those threats.
- Assess the cyber threats against a current system as an input to the risk assessment process.
- Develop their own CSF Organizational Profile that tailors the NCCoE-published *Genomic Data Profile* to identify and prioritize threat-informed mitigations.
- Assess the threat reduction value to the organization of proposed new mitigations.
- Evaluate the risk to a system of a cyber incident or vulnerability that the organization may be concerned about due to recent news events or threat intelligence.

- Assess proposed enhancements to the current system functionality for additional threats that should be considered because of the proposed changes.

While this paper focuses on cybersecurity threats, the NCCoE is developing a privacy-focused guide to address privacy-related concerns, threats, and risks that will also be published.

## References

[1] Genetic Information Nondiscrimination Act of 2008, Pub. L. 110-233, 122 Stat. 811. https://www.govinfo.gov/app/details/PLAW-110publ233

[2] Executive Order 14081 (2022) Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy. (The White House, Washington, D.C.), DCP-202200786, September 12, 2022. https://www.govinfo.gov/content/pkg/FR-2023-04-27/pdf/2023-08841.pdf

[3] Pulivarti R, Martin N, Byers F, Wagner J, Maragh S, Wilson K, Wojtyniak M, Kreider B, Frances A, Edwards S, Morris T, Sheldon J, Ross S, Whitlow P (2023) Cybersecurity of Genomic Data. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency or Internal Report (IR) NIST IR 8432. https://doi.org/10.6028//NIST.IR.8432

[4] Martin N, et al. (2023) Cybersecurity Framework Profile for Genomic Data. (National Institute of Standards and Technology, Gaithersburg, MD), Initial Public Draft NIST Interagency or Internal Report (IR) 8467. https://doi.org/10.6028/NIST.IR.8467.ipd

[5] National Institute of Standards and Technology (2006) Minimum Security Requirements for Federal Information and Information Systems. (Department of Commerce, Washington, D.C.), Federal Information Processing Standards Publications (FIPS PUBS) 200. https://doi.org/10.6028/NIST.FIPS.200

[6] Dodson DF, Montgomery DC, Polk WT, Ranganathan M, Souppaya MP, Johnson S, Kadam A, Pratt C, Thakore D, Walker M, Lear E, Weis B, Barker WC, Coclin D, Hojjati A, Wilson C, Jones T, Baykal A, Cohen D, Yeich K, Fashina Y, Grayeli P, Harrington J, Klosterman J, Mulugeta B, Symington S, Singh J (2021) Securing Small-Business and Home Internet of Things (IoT) Devices: Mitigating Network-Based Attacks Using Manufacturer Usage Description (MUD). (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 1800-15. https://doi.org/10.6028/NIST.SP.1800-15

[7] Dempsey KL, Chawla NS, Johnson LA, Johnston R, Jones AC, Orebaugh AD, Scholl MA, Stine KM (2011) Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-137. https://doi.org/10.6028/NIST.SP.800-137

[8] U.S. Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH) (2023) Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions. Available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions

[9] The MITRE Corporation and Medical Device Innovation Consortium (MDIC) (2021) Playbook for Threat Modeling Medical Devices. Available at https://www.mitre.org/sites/default/files/2021-11/Playbook-for-Threat-Modeling-Medical-Devices.pdf

[10] Threat Modeling Manifesto Working Group (2021) The Threat Modeling Manifesto. Available at https://www.threatmodelingmanifesto.org/

[11] Shostack A (2014) Threat Modeling: Designing for Security (Wiley, Hoboken, NJ) 1st Ed. https://www.wiley.com/en-us/Threat+Modeling%3A+Designing+Security-p-9781118809990

[12] Cooper DA, Apon D, Dang QH, Davidson MS, Dworkin MJ, Miller CA (2020) Recommendation for Stateful Hash-Based Signature Schemes. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-208. https://doi.org/10.6028/NIST.SP.800-208

[13] U.S. Food and Drug Administration (FDA) (2024) Medical Device Product Classification of High Throughput DNA Sequence Analyzer. Available at https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=PFF

[14] Saadia A, et al. (2021) Analysis of security and privacy challenges for DNA-genomics applications and databases. *Journal of Biomedical* 119(103815). https://doi.org/10.1016/j.jbi.2021.103815

[15] Ney P, et al. (2017) Computer Security, Privacy, and DNA Sequencing: Compromising Computers with Synthesized DNA, Privacy Leaks, and More. *2017 USENIX Security* Symposium, (University of Washington, Seattle, WA) pp. 1-15. https://dnasec.cs.washington.edu/dna-sequencing-security/dnasec.pdf

[16] Chard K, et al. (2018) The Modern Research Data Portal: a design pattern for networks, data-intensive science. *PeerJ Computer Science* 4:e144. https://doi.org/10.7717/peerj-cs.144

[17] Chandramouli R (2022) Guide to a Secure Enterprise Network Landscape. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-215. https://doi.org/10.6028/NIST.SP.800-215

[18] National Security Agency (NSA) and Cybersecurity Infrastructure & Infrastructure Security Agency (CISA) (2021) Selecting a Protective DNS Service, Ver. 1.2. Available at https://media.defense.gov/2021/Mar/03/2002593055/-1/-1/0/CSI_PROTECTIVE%DNS_UOO117652-21.PDF

[19] PacBio (2024) Revio™ system: Site preparation guide. Available at https://www.pacb.com/wp-content/uploads/Guide-Revio-system-site-preparation.pdf.

[20] PacBio (2024) Featured Documentation. Available at https://www.pacb.com/support/documentation/

[21] OpenSSL enc - symmetric cipher routines. Available at https://docs.openssl.org/1.1.1/man1/enc/

[22] Quobyte (2024) Quobyte. Available at https://www.quobyte.com

[23] Grace T, Nolan T, Burke K, Dudley R, White G, Good T (2006) Guide to Test, Training, and Exercise Programs for IT Plans and Capabilities. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-84. https://doi.org/10.6028/NIST.SP.800-84

# Appendix A. Abbreviations and Acronyms

The following acronyms are used in this publication.

**ACL**
Access Control List

**API**
Application Programming Interface

**AWS**
Amazon Web Services

**BAM**
Binary Alignment Map

**CRAM**
Compressed Reference-oriented Alignment Map

**CRISPR**
Clustered Regularly Interspaced Short Palindromic Repeats

**CSF**
NIST Cybersecurity Framework

**DFD**
Data Flow Diagram

**DISA**
Defense Information Systems Agency

**DMZ**
Demilitarized Zone

**DNA**
Deoxyribonucleic acid

**DNS**
Domain Name System

**DoS**
Denial of Service

**EDL**
External Dynamic List

**EoP**
Elevation of Privilege

**FDA**
U.S. Food and Drug Administration

**GB**
Gigabytes

**HTTPS**
Hypertext Transfer Protocol Secure

**HVD**
High Value Dataflow

**ICS**
Industrial Control Systems

**ID**
Identifier

**IP**
Internet Protocol

**IT**
Information Technology

**LDAP**
Lightweight Directory Access Protocol

**LIMS**
Laboratory Information Management System

**MDIC**
Medical Device Innovation Consortium

**MFA**
Multifactor Authentication

**MO**
Mission Objective

**NCBI**
National Center for Biotechnology Information

**NCCoE**
National Cybersecurity Center of Excellence

**NFS**
Network File System

**NIH**
National Institutes of Health

**NIST**
National Institute of Standards and Technology

**NTP**
Network Time Protocol

**OS**
Operating System

**OSS**
Open-Source Software

**PaaS**
Platform as a Service

**Playbook**
*Playbook for Threat Modeling Medical Devices* (MITRE and Medical Device Innovation Consortium, 2021)

**QC**
Quality Control

**RBAC**
Role-Based Access Control

**REST**
Representational State Transfer

**RNA**
Ribonucleic Acid

**SaaS**
Software as a Service

**SMB**
Server Message Block

**SP**
NIST Special Publication

**SSDF**
Secure Software Development Framework

**SSO**
Single Sign-On

**STIGs**
Security Technical Implementation Guides

**STRIDE**
Spoofing, Tampering, Repudiation, Information Disclosure, and Elevation of Privilege threat analysis

**TCP**
Transmission Control Protocol

**TLS**
Transport Layer Security

**TTPs**
Tactics, Techniques, and Procedures

**UI**
User Interface

**VM**
Virtual Machine

**VPN**
Virtual Private Network