NIST Internal Report
NIST IR 8432 ipd

Cybersecurity of Genomic Data

Initial Public Draft

Ron Pulivarti
Natalia Martin
Fred Byers
Justin Wagner
Justin Zook
Samantha Maragh
Kevin Wilson
Martin Wojtyniak
Brett Kreider
Ann-Marie France
Sallie Edwards
Tommy Morris
Jared Sheldon
Scott Ross
Phillip Whitlow

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8432.ipd

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8432 ipd

Cybersecurity of Genomic Data

Initial Public Draft

Ron Pulivarti
Natalia Martin
Fred Byers
National Cybersecurity Center of Excellence
Information Technology Laboratory

Justin Wagner
Justin Zook
Samantha Maragh
Material Measurement Laboratory

Kevin Wilson
Martin Wojtyniak
Brett Kreider
Ann-Marie France
Sallie Edwards
MITRE

Tommy Morris
Jared Shelton
University of Alabama in Huntsville

Scott Ross
Phillip Whitlow
HudsonAlpha

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8432.ipd

March 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at https://csrc.nist.gov/publications.

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

How to Cite this NIST Technical Series Publication:
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 ipd. https://doi.org/10.6028//NIST.IR.8432.ipd

Author ORCID iDs [will be updated in the final publication]
Ron Pulivarti: 0000-0002-8330-3474
Natalia Martin: 0000-0000-0000-0000
Fred Byers: 0000-0000-0000-0000
Justin Wagner: 0000-0000-0000-0000
Justin Zook: 0000-0003-2309-8402
Samantha Maragh: 0000-0003-2564-9589

Public Comment Period
March 3, 2023 - April 3, 2023

Submit Comments
genomic_cybersecurity_nccoe@nist.gov

National Institute of Standards and Technology
Attn: Applied Cybersecurity Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 2002) Gaithersburg, MD 20899-8930

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

Abstract

Genomic data has enabled the rapid growth of the U.S. bioeconomy and is valuable to the individual, industry, and government because it has multiple intrinsic properties that in combination make it different from other types of high value data which possess only a subset of these properties. The characteristics of genomic data compared to other high value datasets raises some correspondingly unique cybersecurity and privacy challenges that are inadequately addressed with current policies, guidance documents, and technical controls.

This report describes current practices in risk management, cybersecurity, and privacy management for protecting genomic data along with relevant challenges and concerns. Gaps in protection practices across the lifecycle were identified concerning genomic data generation, safe and responsible sharing of the genomic data, monitoring the systems processing genomic data, lack of specific guidance documents addressing the unique needs of genomic data processors, and regulatory/policy gaps with respect to national security and privacy threats in the collection, storage, sharing, and aggregation of human genomic data.

The report proposes a set of solution ideas that address real-life use cases occurring at various stages of the genomic data lifecycle along with candidate mitigation strategies and the expected benefits of the solutions. Additionally, areas needing regulatory/policy enactment or further research are highlighted.

Keywords

Cyberbiosecurity; cybersecurity; genomic data; human genome; genomics; privacy.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems.

Additional Information

For additional information on this NIST Internal Report and the NCCoE Genomics Cybersecurity project, visit our project page. Information on other efforts at NIST, the Information Technology Laboratory and NCCoE is also available.

Call for Patent Claims

This public review includes a call for information on essential patent claims (claims whose use would be required for compliance with the guidance or requirements in this Information Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may be directly stated in this ITL Publication or by reference to another publication. This call also includes disclosure, where known, of the existence of pending U.S. or foreign patent applications relating to this ITL draft publication and of any relevant unexpired U.S. or foreign patents.

ITL may require from the patent holder, or a party authorized to make assurances on its behalf, in written or electronic form, either:

a) assurance in the form of a general disclaimer to the effect that such party does not hold and does not currently intend holding any essential patent claim(s); or

b) assurance that a license to such essential patent claim(s) will be made available to applicants desiring to utilize the license for the purpose of complying with the guidance or requirements in this ITL draft publication either:

   i. under reasonable terms and conditions that are demonstrably free of any unfair discrimination; or

   ii. without compensation and under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

Such assurance shall indicate that the patent holder (or third party authorized to make assurances on its behalf) will include in any documents transferring ownership of patents subject to the assurance, provisions sufficient to ensure that the commitments in the assurance are binding on the transferee, and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding each successor-in-interest.

The assurance shall also indicate that it is intended to be binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

Such statements should be addressed to: genomic_cybersecurity_nccoe@nist.gov

Table of Contents

List of Tables

List of Figures

Acknowledgments

The authors would like to thank the speakers, panelists, and contributors to the NCCoE Virtual Workshop on the Cybersecurity of Genomic Data held Wednesday, January 26, 2022, and the NCCoE Virtual Workshop on Exploring Solutions for the Cybersecurity of Genomic Data held Wednesday, May 18, 2022, and Thursday, May 19, 2022, for their inputs, which helped inform the contents of this paper.

1. Introduction

The world has entered an era of accelerated biological innovation built primarily upon the many uses of genomic data that include vaccine development and manufacturing, pharmaceutical development and manufacturing, disease diagnosis, agricultural innovations that enable increased food production, biofuel development, basic and translational scientific research, consumer testing, genealogy, and law enforcement, among others. More uses continue to be discovered. Genetic sequencing technology has advanced such that sequencing entire genomes is feasible and affordable. Whole or partial genome sequences for many microbial, plant, and animal species reside in open access, controlled access, or private databases within the National Institutes of Health (NIH), Federal Bureau of Investigation (FBI), and direct-to-consumer (DTC) genetic testing providers, to name a few. As this era unfolds there is a new awareness of risks to U.S. national security, its economy, its biotechnology industry, and its citizens due to cybersecurity attacks targeting genomic data as highlighted in the Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy [1]. Additionally, human genetic information requires complying with policies, laws, and ethics surrounding privacy. Nevertheless, the inherent value of some genomic data lies in the ability to share information with the broader community, creating the need to balance access restrictions with data sharing capabilities.

1.1. Cybersecurity and Privacy Concerns[1]

Cyber attacks targeted at genomic data include attacks against the confidentiality of the data, its integrity, and its availability. Cyber attacks against the confidentiality of the data can threaten our economy through theft of the intellectual property owned by the U.S. biotechnology industry, allowing competitors to gain an unfair economic advantage by accessing U.S. held genomic data. Attacks against the integrity of the data can disrupt biopharmaceutical output, agricultural food production, and bio-manufacturing activity. Attacks against the availability of the data include encrypting for ransom, deletion of data, and disabling critical automated equipment used in research, development, and manufacturing. The potential harms of cyber attacks on genomic data threaten our national security as well, including enabling the development of biological weapons and the surveillance, oppression, and extortion of our citizens, military, and intelligence personnel based on their genomic data.

Cyber attacks targeted at genomic data can also harm individuals by enabling blackmail, discrimination based on disease risk, and privacy loss from the revealing of hidden consanguinity or phenotypes including health, emotional stability, mental capacity, appearance, and physical abilities. In addition to the privacy risks that can arise because of a cyber attack, privacy risks unrelated to cybersecurity can arise when processing genomic data. These risks can arise when

[1] Cybersecurity and privacy objectives provide a useful construct for describing concerns. The cybersecurity objectives are confidentiality, integrity, and availability. The definitions of each originate from 44 U.S.C., Sec. 3542 and they are used throughout multiple NIST publications. The privacy engineering objectives are predictability, manageability, and disassociability. They were initially described in NISTIR 8062 and are also used throughout multiple NIST publications. Definitions for all six terms, as well as a sample of documents in which they are discussed, appear in the NIST Glossary (available at: https://csrc.nist.gov/glossary).

there is insufficient predictability, manageability, and disassociability in the genomic data processing. Insufficient predictability in data processing can result in privacy problems if individuals are surprised by what is happening with their genomic data. Insufficient manageability in data processing can arise when the capabilities are not in place to allow for appropriately granular administration of genomic data, for example, individuals may need to be able to have some or all their genomic data deleted from a dataset. Permitting access to raw genomic data, instead of using appropriate privacy-enhancing technologies to extract only the necessary insights (without revealing the raw data), introduces privacy risks from insufficient disassociability in data processing. Each of these areas of privacy risks can disrupt the ability to realize the benefits of processing genomic data [2].

NIST is exploring genomic data uses to better understand common and pressing cybersecurity and privacy concerns specific to this data to identify and provide security and privacy practice guidance to help protect it. To inform this effort, NIST, through its National Cybersecurity Center of Excellence (NCCoE), conducted two public workshops [3][4]—one concentrating on the challenges already faced or anticipated by the community, followed by another workshop focusing on solutions to help address those challenges. These workshops along with additional research provided the basis for the information presented in this paper.

1.2. Document Scope and Goals

This paper identifies characteristics of genomic data compared to other types of high-value data and provides an introductory overview of cybersecurity and privacy risk management resources and classifications of the risks during the genomic data lifecycle. It also identifies the most common challenges to securing genomic data, the current state-of-the-art cybersecurity and privacy practices for genomic data, and gaps associated with these practices. Finally, a set of use cases are presented that simulate real life challenges with candidate mitigation strategies to address each challenge, along with the expected benefits provided by the proposed solutions.

2. Background

A genome contains hereditary material comprised of nucleic acids, mostly in the form of deoxyribonucleic acid (DNA). Genomes contain the full set of instructions to form an organism and are largely unchanged from conception to death. Instructions are encoded in the sequence of the four nucleotide subunits (also called bases) that comprise the backbone of the DNA or Ribonucleic acid (RNA) molecule. In DNA, these are adenine (A), cytosine (C), guanine (G), and thymine (T). A segment of bases containing the instructions for making a product, such as a protein or an RNA molecule, is called a gene. Variations in the number and kinds of genes, as well as differences across genes, underpin the diversity of life on earth.

Some genes give rise to observable traits, termed phenotypes, like hair or eye color, blood type, and facial features. Other phenotypes may include the presence of or susceptibility to certain medical conditions, such as sickle cell anemia, cystic fibrosis, Huntington's disease, forms of muscular dystrophy, or certain cancers. Importantly, the genome reveals a great deal of information about an organism as well as its relatives.

Genomic data is immutable, associative, and conveys important health, phenotype, and personal information about individuals and their kin (past and future). In some cases, small fragments of genomic data stripped of identifiers can be used to re-identify persons, though the vast majority of the genome is shared among individuals [3][4][5][6][7][8].

2.1. Genomic Information Lifecycle

DNA sequencing, a key method researchers use to understand genomic information, identifies the order of base pairs across a stretch of DNA or even the whole genome. This information helps researchers understand genome organization and identify any similarities or differences that may be present across organisms. DNA sequence data are generated by different stakeholders, including academic researchers, healthcare providers, government agencies, and industry. These entities may generate and store sequence data for internal use, make data available for public access, distribute it to select entities or communities of interest, and/or aggregate their results with others.

Figure 1 summarizes the genomic lifecycle and contrasts those that are out of scope for this report: Sample Collection and Sample Preparation with what is considered in scope: Sequencing, Data Generation, and Data Analysis.

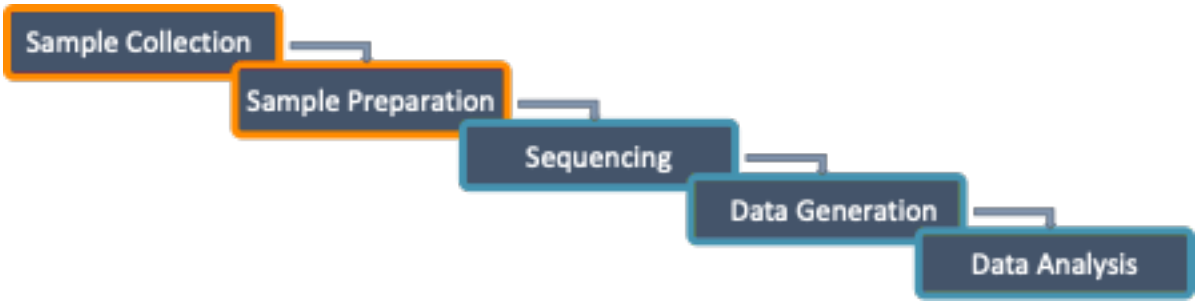


Fig. 1. Genomic Data Lifecycle; Naveed et.al. [6]

The genomic data lifecycle begins with sample collection, followed by DNA extraction and several processing and quality control steps performed in the laboratory. Sample preparation includes the processes where collected DNA is purified, fragmented, and bound to adapter molecules to generate a library for sequencing. Libraries are collections of DNA molecules that are compatible with the sequencing system. The prepared sample is then applied to a specially designed substrate, like a glass flow cell, that is used by the sequencing system to read the order of the DNA bases. Safeguarding sample collection and DNA processing requires applying appropriate physical security as opposed to cybersecurity controls, and is deemed out of scope. One exception is the collection of human DNA in the context of clinical care or study. In those instances, human DNA sample collection and downstream steps are subject to additional privacy considerations and informed consent.

Once a sample has been sequenced, the instrument outputs raw data. Resulting raw data file formats differ across manufacturers of sequencing systems but typically contain identifiers, reads (e.g., DNA sequences read by the instrument), and quality scores for each base call.

The resulting raw data are typically stored in a computer on-premises or in the cloud for processing and analysis. Data processing and analyses can be performed by open-source or commercially developed software or algorithms. The steps performed vary based on the intended use of the sequence data. Example data processing steps include demultiplexing (if multiple samples were sequenced within a single run), filtering out low-quality reads, aligning to a reference genome, identifying variants, or annotating genomic variants (e.g., with how they may change genes). The data may then be analyzed to identify, for example, variants across samples or correlate variants with phenotypes. Common clinical applications of DNA sequencing include diagnosing inherited diseases, targeting more effective drugs with precision medicine (e.g., for cancer), and better understanding genetic risk factors for diseases.

These data may be made accessible to other researchers or uploaded to public or controlled access databases, like the sequence read archive (SRA) or the database of Genotypes and Phenotypes (dbGaP) that is run by the National Center for Biotechnology Information (NCBI) at the NIH.

In healthcare, genomic data can be medically relevant at any time in a patient's life, thus integration with an electronic health record (EHR) is essential. However, the nature and size of the data presents a challenge with incorporating it into EHR workflows. A whole genome sequence can be used to identify an individual in some cases, and the raw data can exceed 100 gigabytes in size. Additionally, as research advances, a genomic sequence may need to be re-analyzed to provide the most up-to-date information for a patient's current medical condition. Thus, the portability, privacy, chain-of-custody, interoperability, consent management, and re-interpretation of genomic data are all key points for its effective use in healthcare.

Like other sensitive data, at each stage in the genomic data lifecycle from creation to storage and analysis to dissemination, the data can be at risk of being intercepted, corrupted, overwritten, or deleted.

2.2. Next Generation Sequencing

Since the introduction of automated sequencers in 1986 [9], sequencing costs have dropped by several orders of magnitude [10]. Both the number and types of DNA sequencers have proliferated for clinical and research applications. The first sequencing platforms used Sanger-based chemistries, relying on the chain termination method, and were limited in the numbers of sequences that could be read simultaneously. Advances in sequencing approaches and chemistries have led to second- and third-generation sequencing platforms, often called next-generation sequencing (NGS). NGS systems have been developed by several manufacturers with varying read lengths and accuracy for different applications. Compared to automated Sanger sequencing, which reads a single DNA molecule many times, NGS systems read many molecules simultaneously, greatly increasing throughput and efficiency while reducing costs.

2.3. Variant Calling

Variant calling, the primary form of genomic analysis, is the identification of differences in the genetic sequence between a sample and a reference genome. Variants can range from changes to a single base at a given position in the genome to larger structural alterations, such as deletions, insertions, duplications, inversions, and translocations. Variant identification is particularly important in clinical medicine and molecular biology as alterations elucidate the role of genetic elements in various diseases and facilitate our understanding of cellular processes.

Characterizing cancers is one area that employs variant identification. For instance, analyses of somatic and germline genetic alterations—germline mutations are inherited by germ cells (sperm and egg) during conception whereas somatic mutations occur after conception in non-germ cells—yield insights to clinical outcomes and tumor origins. Somatic variants can be used to target drugs that are more likely to be effective against the individual's tumor.

As NGS technologies have progressed, so too have the computational pipelines for identifying variants. Multiple software tools and workflows have been developed for different NGS technologies and various variant calling applications. Precise variant calling requires having complete and accurate reference genomes, as well as appropriate sequencing depth, accuracy, read length, and analysis methods. Efforts including but not limited to the NIH Genome Reference Consortium, Telomere-to-Telomere Consortium, and the NIST Genome in a Bottle have significantly advanced the technical infrastructure, accuracy, and completeness of human reference genomes and authoritatively characterize genomes to assess accuracy of variants [11][12]. Recent systematic evaluations of germline and somatic variant calling pipelines showcased the capabilities and limitations of several sequencing platforms and variant calling software and workflows [13][14]. As sequencing systems continue to advance, variant calling pipelines are likely to follow suit.

2.4. Genome Editing

New advances in DNA editing techniques, like clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated protein (CRISPR-Cas), promise precise, efficient, and affordable ways to edit the genome through removing, adding, or altering DNA segments. As

next generation sequencing is helping scientists better understand the genomes of various organisms, CRISPR-Cas systems are enabling researchers to modify genes more easily and affordably than ever possible. When using genome editing in clinical applications to treat diseases, NGS technologies are used to confirm the correct edit was made and identify any unintended off-target edits. Although the net benefit of these technological advances is positive, there exists an opportunity for misuse.

2.5. Direct-to-Consumer Testing

An alternative approach to measure DNA does not use sequencing but instead uses microarrays of DNA probes to measure common variants at millions of specific genomic positions. Microarrays are commonly used by DTC genetics companies for ancestry analysis and screening for whether individuals may be carriers for particular disease-related variants. DTC genetic tests are marketed to individuals without the involvement of a healthcare provider. These tests, available from multiple companies, enable consumers to gain personalized insights into their health and ancestry by examining their genetic profile. Results from consumers' genetic tests can be accumulated in research databases to identify genetic sequence associations with certain conditions. Moreover, third-party online genetic genealogy services allow consumers to upload their genetic sequence information and identify related individuals.

Advancements in microarray technology have allowed these companies to provide tests at relatively affordable price points. The costs of DTC tests vary from company to company but typically start at $99. The DTC genetic testing market is estimated to be worth over $1.3 billion and is projected to grow to approximately $3.5 billion by the end of 2026 [15].

2.6. The Characteristics of Genomic Data

Genomic data share attributes with other sensitive types of information, and as such, mirrors their need for secure storage and transfer. Beyond these aspects there are several intrinsic characteristics of genomic data. These concepts were discussed during the first NCCoE public workshop held on May 18 and 19, 2021, and have been posited by some in the research community [3]. Figure 2 identifies seven features of genomic information that distinguish it from other types of data. It is not any single characteristic, but instead the combination of these intrinsic properties that highlight its value and sensitivity.

- **Phenotype.** Phenotype refers to the observable characteristics imparted by the genome, such as size, appearance, blood type, and color. DNA can reveal a great deal of information about an individual or their relatives, including phenotype or health information.
- **Health.** Health means that DNA contains information about an organism's disease presence, disease risk, vigor, and longevity. Clinical genetic testing can identify variants within one's genome that may contribute to certain health outcomes.
- **Immutable.** Immutable means that an organism's DNA does not change significantly during the organism's life. An individual's genome is practically immutable, with a

negligible lifetime mutation rate for most applications, which increases the long-term consequences of a data breach.

- **Unique.** Unique means that individuals of species with sexual reproduction can be identified. Except in the case of identical siblings, a person's genome is unique to them.
- **Mystique.** Mystique refers to the public perception about the mystery of DNA and its possibly future uses.
- **Value.** Value refers to the importance of the information content of DNA that does not decline with time, but typically increases with time. Genomic information has value to certain entities and that value is predicted to grow as we learn more about the genomes of humans and other organisms.
- **Kinship.** Kinship means that common ancestors and descendants of the organism can be identified from DNA samples. Consumer genetic testing services provide information about one's ancestral lineage, including the potential to identify relatives.

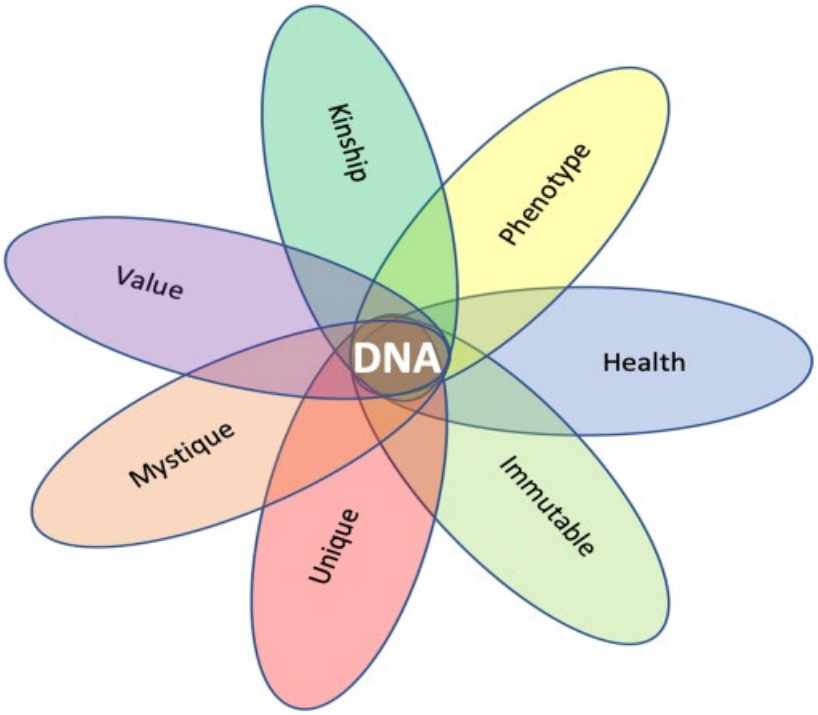


Fig. 2. Characteristics of DNA (Adapted from Figure 1 of Naveed et al. [6])

2.7. Balance Between Benefits and Risks for Uses of Genomic Information

The U.S. research community, government, and private industry require genomic data sharing to advance scientific and medical research and to maintain the country's competitive advantage in biotechnology. The transfer and sharing of genomic data are essential for understanding human health, improving wellbeing, and accelerating scientific inquiry and advancements. For example, in 2021 the NIH processed almost 40,000 requests for data access and has about three million genotype microarray datasets and over 500,000 whole genome sequences [4][15]. Genomic data

enables precision medicine for rare diseases and cancer and is an enabler for CRISPR technology. DTC genomic testing allows people to benefit from ancestry tracing, relative matching, and health insights. In 2022, DTC companies reported more than 40 million consumers have used their services [17]. Genomic data can be useful for forensics to solve crimes [18]. Future uses of genomic data have further potential to advance the bioeconomy.

The genomic data transferred and shared represents tens of millions of individuals who provide their information. In aggregate across all types of measurements, this data is touched by thousands of entities (e.g., domestic, international, nonprofit, for-profit) that store, access, manage, and use genomic and health-related data. These data sharing activities need adequate technological and policy controls that allow research and enable commerce, as well as respect the informed consent and privacy of the data subjects who expect protections from re-identification [19].

Loss of control of genomic data can cause risks to privacy, personal security, and national security, as adversaries can use genomic data for nefarious reasons such as surveillance, oppression, and extortion. Genomic database breaches or other losses of data may result in thefts of intellectual property and put the U.S. at a competitive disadvantage in biotechnology. As reported by national security experts, security threats may arise through the creation of population-specific bioweapons or compromised identities of national security agents [3][20]. Cyber attacks have occurred on genomic databases, DNA sequencing instruments, and genomic software tools [21][22][23][24][25].

3. Challenges and Concerns Associated with Handling Genomic Information

This section summarizes challenges and concerns for protecting genomic data in national security, personal security and privacy, discrimination and reputational, economic, health outcomes, and other potential future concerns. These challenges and concerns come from discussions and presentations from the workshops and through literature review.

3.1. Potential National Security Concerns

As genomics research evolves, threats to national security continue to emerge. This section lists potential national security concerns that were identified by workshop stakeholders. The likelihood for these threats was not evaluated as part of the workshop.

The national security concerns identified include:

- **Infringement:** Genomic data can be used for population surveillance and oppression a well as extortion of our citizens, military, and intelligence personnel [21].
- **Biological weapons:** Potential risks related to bioweapons directed against an individual's DNA sequence or cloning an individual were deemed to be impractical and not concerns that should drive genomics cybersecurity work [26].
- **Production of toxic products or infectious agents:** Peccoud J., et.al. [27], speculates that due to the lack of integrity and tampering controls that typically exist, genomic data could be corrupted by altering sequences or annotations and that, "These changes could delay research programs or result in the uncontrolled production of toxic products or infectious agents." This could result in significant loss of life as well as economic consequences.

3.2. Privacy Challenges

Privacy challenges resulting from the use of human genomic data include problems for individuals such as enabling blackmail, discrimination based on disease risk, and the revelation of hidden consanguinity or phenotypes including health, emotional stability, mental capacity, appearance, and physical abilities.

The U.S. intelligence community highlighted [3][21] concerns with the sparse regulation of the genomic data collected and stored by DTC companies, along with the lack of recognition of U.S. genomic data as an asset that needs export controls. Some researchers and consumer groups suggested that due to the high reidentification risk, human genomic data should be consistently classified as personally identifiable information (PII), while other researchers argued that this would hinder research and the benefit to society [3]. It should be noted that regardless of the classification, the privacy risks arising from processing genomic data remain.

Potential privacy problems identified include:

- **Re-identification of de-identified genomic data:** Human genomic data, even small fragments of a person's whole genome, can usually be re-identified for some populations when combined with available datasets, such as ancestry data, self-shared identified genomic data of distant relatives, surname inference, age, etc. [6][7][8][28].

- **Unanticipated revelation of individuals' blood-relatives can lead to dignity loss when those relationships are identified:** Consanguineal ties may be revealed that may be embarrassing or incriminating, resulting in psychological or reputational harm.

3.3. Discrimination and Reputational Concerns

The discrimination and reputational concerns identified include:

- **False identification:** Sample mishandling, crime lab procedure deviations, or intentional modification of the digital genomic data produce a risk to individuals being framed for crimes they did not commit or extorted with the falsified genomic information.
- **Discrimination based on disease risk identified by an individual's genomic data:** While some forms of discrimination in the U.S. are prohibited by law, the federal laws are narrowly written and do not prohibit discrimination based on an individual's genomic data for things such as life insurance, acceptance into the military, or senior residential communities (e.g., based on Alzheimer's risk) [6][29].
- **Unintended consequences from sample bias**: Artificial Intelligence (AI) and statistical analysis techniques analyze large sets of genomic data to find diseases, risk factors for diseases, make treatment decisions, and predict patient prognosis. Large datasets from DTC and other samples of convenience in the U.S. are from predominantly those of European descent. Minority communities are typically underrepresented. This bias of the sample data can be amplified, particularly in AI techniques, and impact the results of these analysis and prediction techniques that can result in discrimination [29][30][31], resulting in potential harms to those whose genomic data are not represented in the sample set. This bias of AI algorithms has been studied by NIST in the field of facial recognition [32].

3.4. Economic Concerns

Economic concerns identified include:

- **Intellectual property infringement:** Exfiltration of genomic data can result in loss of intellectual property for institutions, resulting in economic losses in the future.
- **Operational disruption:** Disruptions to the generation, storage, or use of genomic data can negatively affect production operations, impacting agricultural food production, biopharmaceutical output, and biomanufacturing.
- **Extortion:** Bad actors or nation states may extort individuals based on their genomic data, threatening to reveal sensitive health or kin information encoded in the individual's genome [21], resulting in financial, psychological harm, and reputational loss to the individual.
- **Penalties and liabilities:** Negligence in the cybersecurity controls of genomic data by the company or institution could violate their regulatory obligations, jeopardize their access to data in the future, and result in significant financial loss through imposed penalties.
- **Revoke data access:** Negligence in adequate security or privacy protection can cause individuals to no longer consent to have their genomic data used in scientific studies, which could reduce the effectiveness of research and threaten the bioeconomy.

3.5. Health Outcome Concerns

A patient's genomic data needs to be effectively incorporated into their EHR to allow for effective clinical care. Concerns in healthcare include portability, chain-of-custody, re-interpretation of genomic data, and consent management.

3.6. Other Potential Future Concerns

Other potential future concerns identified include:

- **Loss of individuality:** While not a realistic threat at present or in the near future, cloning of an individual could theoretically be accomplished with their genomic information. This could result in psychological or financial harm to the person and their clone [33].
- **Human engineering:** Genomic information could be used for eugenics, more finely targeting kinship or other human traits [34]. This could result in societal harm or psychological harm to the individual.

3.7. Summary of Challenges and Concerns with Genomic Data

The U.S. research community, government, healthcare, and private industries handle genomic data and require genomic data sharing to advance scientific and medical research, improve health outcomes, and maintain the country's competitive advantage in biotechnology. Genomic data often needs to be aggregated from multiple studies to address pressing research questions, but challenges such as the differing subject consents used in the different studies can present difficulties that need adequate technological and policy controls that respect the informed consent and privacy of the data subjects [3]. Though it can be time-intensive and difficult, responsible data sharing and analytics facilitates commerce, research, and healthcare outcomes while protecting subject privacy against re-identification and respecting subject informed consent. Further, additional guidance is needed to promote the safe and secure sharing of genomic data while addressing threats to national security, personal security and privacy, reputations and civil liberties, the economy, health, and the future of humanity.

4. Current State of Practices

This section identifies many of the cybersecurity, privacy, and risk management practices issued by U.S. Government, industry, and international entities. While cybersecurity and privacy practices are inter-related and support overall risk management practices, this section identifies practices specific to each area. Subsections highlight the NIST Risk Management Framework (RMF), the NIST Cybersecurity Framework, and the NIST Privacy Framework, along with legislation, frameworks, alliances, and other resources that can be leveraged to improve the protection of genomic data. The final subsection identifies guidance, technical, and policy/regulatory gaps that can then be addressed by solutions proposed in the subsequent section.

4.1. Risk Management Practices

Risk management is the process of managing risks to organizational assets and operations, individuals, and other entities (e.g., nations). The NIST RMF [35] provides a well-established method for information security risk management that can apply to any system or organization. Federal risk management related initiatives such as Executive Order (EO) 14028 [36] and the subsequent guidance published by NIST define methods for protecting the software supply chain. The "Framework for Responsible Sharing of Genomic and Health-Related Data" [37] provides an additional resource from the international community.

4.1.1. U.S. Government Resources

For the U.S. Federal government, the Federal Information Security Management Act (FISMA 2002) served as the initial driver for cybersecurity risk management programs. The Office of Management and Budget (OMB) Circular A-130, "Managing Federal Information as a Strategic Resource" requires Executive agencies to leverage NIST guidance. The NIST RMF and the Federal Risk and Authorization Management Program (FedRAMP) are examples of risk management processes used by federal agencies.

FISMA required each federal agency to develop, document, and implement an agency-wide program to provide information security for the information and systems that support the operations and assets of the agency, including those provided or managed by another agency, contractor, or other sources. The Federal Information Security Modernization Act (FISMA 2014) [38] includes updates to address evolving cybersecurity concerns, reduce reporting burdens, strengthen continuous monitoring in systems, and reporting of incidents.

OMB Circular A-130 requires executive agencies within the federal government to plan for security, ensure that appropriate officials are assigned security responsibility, periodically review the security safeguards in their systems, and authorize system processing prior to operations and periodically, based on risk.

The NIST RMF (defined in NIST SP 800-37 Revision 2) [35] provides a structured, yet flexible, process for managing cybersecurity and privacy risk that includes steps for preparation, system categorization, control selection, control implementation, control assessment, system authorization, and continuous monitoring. Risk management involves more than complying with

613 regulations or technical controls and should be tailored to each organization's mission,
614 regulatory environment, and risk tolerance. Table 1 NIST Risk Management Framework
615 Discussion for Genomic Data provides an overview of the RMF and a brief description of its
616 relevance to genomic data.

617 **Table 1.** NIST Risk Management Framework Discussion for Genomic Data

RMF Step	Overview	Genomic Data Relevance
Prepare	The organization carries out essential activities to prepare for risk management. This includes identifying key risk management roles, establishing an organization-wide risk management strategy and risk tolerance, assessing organization-wide security and privacy risks, and developing an organization-wide continuous monitoring strategy. At the system level, the organization must understand information types in the system, conduct a system-level risk assessment, identify the relevant requirements—including the applicable federal and state regulatory requirements.	It is important that all the relevant regulatory requirements are considered at a state or national level depending on the subject's citizenship, where the data was collected, and where it is being processed. An important feature of human genomic and health data that organizations must prepare for is that the informed consent may have specific restrictions on how that data is used and that the informed consent often differs depending on when and under what circumstances the data was collected.
Categorize	Using outputs of the Prepare step, categorize the system and information processed, stored, and transmitted and gauge the potential loss of the Confidentiality, Integrity, and Availability (CIA) triad for the information, system, and processes.	Data categorization helps an organization identify the appropriate controls to properly mitigate the risk to an acceptable level. Human health and genomic data may additionally be sub-categorized by the allowed uses of the subject's informed consent which may restrict how it is processed.
Select	Controls are selected to manage risk in accordance with the risk management strategy and within the risk tolerance levels.	Specific controls related to genomic data have yet to be identified but could be used as guidance for organizations managing related risks.
Implement, Assess, and Authorize	The system security plans are updated to reflect the implemented state of the controls. Controls are assessed (which includes validation and verification) to ensure they are in place, operating as intended, and achieving the desired results. The Authorize step requires a senior official to understand the residual risks and agree that they are acceptable to the organization before the system is put into operation (or continues to operate).	These steps enable the organization to quantify and characterize the risks associated with the current protections implemented (or not implemented) for the genomic data. Organizations would then manage inherent risk or residual risk through plans of action and milestones (POA&Ms) identifying the resources and timelines required to address residual risks.
Continuous Monitoring	Once the system is operational, an organization must ensure the system is operating as intended and within the acceptable risk tolerance of the organization. Subject matter experts with appropriate expertise are necessary at this stage. Periodic evaluation is needed to keep up with changes as they happen, including new threats, regulatory changes, and technology changes. Organizations must also consider end-of-life procedures for their systems and data.	Organizations must monitor both the relevant threats and outstanding vulnerabilities to determine the risk posed to the organization. Response and recovery plans must be in place to appropriately address events and incidents as they occur to minimize exposure, protect the genomic data, inform users of breaches, and recover from incidents.

FedRAMP [39] is a risk-based approach, aligned to the NIST RMF, for protecting cloud-based services and systems and includes continuous monitoring and an independent assessment requirement. FedRAMP is governed by the OMB, the General Services Administration FedRAMP Program Management Office (PMO), and a FedRAMP Joint Authorization Board. FedRAMP outlines guidance for securing a cloud-hosted system and FedRAMP authorization may be re-used across multiple government agencies. Federal agencies can use a FedRAMP system as part of their overall solution for implementing FISMA requirements.

4.1.2. U.S. Government Initiatives

President Biden's administration issued EO 14028, Improving the Nation's Cybersecurity [36] on May 12, 2021, to direct federal agencies to enhance cybersecurity risk management practices. In response to the EO, NIST issued guidance on critical software and updated guidance on protecting the software supply chain [40]. Genomic sequencing environments can leverage this guidance to implement supply chain protections that will improve visibility into provenance and related software components. More recently, EO 14081, Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy [1] requires identifying cybersecurity risks in biotech.

NIH continues to play a leading role in genomic research and the protection of genomic data. Examples of NIH-led efforts include:

- NIH manages dbGaP, with almost 40,000 requests for access to datasets within this database (2022) that must be shared responsibly.
- NIH is leading genomic-related efforts to move from identity-based access to authority-based access. NIH recommends moving to authority-based access, which focuses on what a researcher and the software are authorized to do within a system, rather than the identity of the researcher. In continually monitoring changing threats, regulations, and technology, NIH has determined that identity-based security models have become insufficient to adequately mitigate the risk for their mission. Identity-based security models lack context and have poor federation. Once a user has been authenticated, there are no further limitations on what the user (or the software used by the user) is authorized to do with the data. Authority-based access uses tokens and is more consistent with the Zero Trust Architecture (ZTA) model and the principle of least privilege [41][42].
- The NIH has been working with the Global Alliance for Genomic Health (GA4GH) to ensure that their implementation of authority-based access is compatible with the GA4GH Passport for researcher authorization.
- NIH has a significant concern about the confinement problem for genomic data sharing—preventing an authorized user from sharing data with others—and continues to look for solutions to this issue. The NIH addresses the confinement problem with contractual controls, as the currently available technical controls have limitations that are not compatible with the NIH mission. However, technical controls that address the confinement problem and have limitations that are more compatible with the NIH mission are being researched.

4.1.3. International Resources and Regulations

The "Framework for Responsible Sharing of Genomic and Health-Related Data" [37] is an international framework specific to genomic data risk management. It addresses international data sharing, collaboration and good governance concerns, and the protection and promotion of the welfare, rights, and interests of individuals from around the world in genomic and health-related data sharing.

In the European Economic Area (EEA), organizations must consider the "Schrems II" ruling [43][44] which addresses confidentiality and data residency of data provided by citizens in the member states of the EEA. In Asia, India is considering expanding its data protection law in a manner similar to the European Union General Data Protection Regulation [45][46], and China is compelling Chinese companies to share data they have collected with the government [21]. Further, China and Russia severely restrict the sharing of genomic information [47][48][49][50].

4.2. Cybersecurity Best Practices

Cybersecurity practices support overall risk management. This section identifies federal, international, and industry resources that organizations can use to help identify, prioritize, and implement cybersecurity capabilities. The NIST Cybersecurity Framework outlines categories of capabilities and provides an overall methodology for prioritizing cybersecurity investments. Another widely used federal resource is the guidance on implementing a ZTA. Additional practices introduced in this section include information sharing and analysis centers, software supply chain management, and the GA4GH Data Security Infrastructure Policy [51].

4.2.1. U.S. Government Resources

The U.S. Government publishes multiple resources that support cybersecurity practices. This section describes 1.) the NIST Cybersecurity Framework, 2.) ZTA guidance, and 3.) Benchmarks or Security Technical Implementation Guides (STIGs), produced by the Defense Information Systems Agency (DISA).

The NIST Cybersecurity Framework [52] is a voluntary, comprehensive framework developed by the federal government, that is applicable to any organization (federal, academic, private sector, etc.) wanting to improve their cybersecurity risk management. It is very broad, not genomic specific, and can be tailored for real-world deployments within specific industries.

ZTA is an evolving cybersecurity paradigm described in NIST SP 800-207 [53] that identifies target requirements and capabilities that align with genomic data protection goals. ZTA enables secure authorized access to resources individually in situations when many users need access from anywhere, at any time, from any device to support the organization's mission. Data is programmatically stored, transmitted, and processed across different boundaries under the control of different organizations to meet ever-evolving business use cases. In a ZTA, no implicit trust is granted to assets or user accounts solely on their physical or network location. Allowing assets and users to only access the required resources for fulfillment of their role in the organization's mission provides for defense-in-depth.

DISA STIGs [54] and similar benchmarks can define secure configurations for cyber-physical systems with operating systems (such as sequencers) that can be assessed and maintained. These guides provide hardening and defense in depth, greatly improving an organization's security posture as the default settings of operating systems are typically optimized for usability and have many insecure settings not appropriate for securing high-value data. Security benchmarks support the secure configuration of hardware and software throughout the genomic lifecycle.

4.2.2. International Resources

The **Bioeconomy Information Sharing and Analysis Center (BIO-ISAC)** [55], an international non-profit organization, addresses threats unique to the bioeconomy, sharing threat intelligence among the stakeholder community. BIO-ISAC facilitates the responsible disclosure of detected or suspected vulnerabilities found in sequencers. This collaborative approach to cybersecurity complements individual organizational efforts. BIO-ISAC is an example of an industry stakeholder providing cybersecurity services, such as emergency threat hunting for organizations who are under active attack, to bioeconomy organizations who choose to augment their own resources.

The GA4GH has developed a **Data Security Infrastructure Policy** [56][57] that is a "set of recommendations and best practices to enable a secure data sharing and processing ecosystem." They provide frameworks and standards for responsible genomic data sharing. While the proposed frameworks and standards provide foundational principles for genomic data sharing, they are high-level guidelines and lack articulation of in-depth and comprehensive security features and solutions that are necessary for a real-world development and deployment of such principles.

4.2.3. Industry Resources

Industry cybersecurity practices including software bill of materials (SBOM), automated vulnerability scanners, and the use of manual expert analysis support the protection of genomic data. These industry practices are becoming more widely implemented even in the U.S. federal government and should be considered for all systems processing genomic data.

SBOMs are widely used across industry and now being promoted by the Cybersecurity and Infrastructure Security Agency (CISA) [58] as an essential building block in software security and software supply chain risk management. SBOMs contain a nested inventory identifying the ingredients that make up software components and provide transparency to cybersecurity professionals and users to facilitate effective patch management and mitigation of newly discovered software vulnerabilities.

Vulnerability scanners are automated tools that examine software systems for misconfigurations and software flaws that compromise the cybersecurity of the system. They are highly useful in detecting unintended mistakes in threat mitigation and are an important part of assessing a system for cybersecurity. Additionally, ZTA guidance from OMB M-22-09 requires federal agencies to conduct manual expert analysis of systems by approved external third-party testing capabilities.

4.3. Privacy Best Practices

This section describes the current state of privacy practices related to genomic data including U.S. Government and industry resources along with relevant regulations.

4.3.1. U.S. Government Resources

U.S. Government privacy resources include the NIST Privacy Framework, Federal laws including Genetic Information Nondiscrimination Act of 2008 (GINA) [59], Health Insurance Portability and Accountability Act (HIPAA) [60] and the Common Rule.

The NIST Privacy Framework [2] is a voluntary framework that helps organizations identify and manage privacy risk within an organization's broader enterprise risk portfolio. Like the NIST Cybersecurity Framework, the NIST Privacy Framework can be used across federal and non-government organizations. While not specific to any use case, it can be tailored to address genomic data privacy requirements for those producing, storing, or processing genomic data.

GINA [59] includes two key provisions that prohibit group insurance providers and employers of more than 15 people from using genetic information to discriminate against individuals.

HIPAA [60] protects patient confidentiality by placing restrictions on sharing protected health information (PHI) by HIPAA-covered entities (i.e., insurance companies and healthcare providers) [61]. A 2013 amendment modified the HIPAA Privacy Rule to consider genetic information as PHI [61]. Non-covered entities, like employers, law enforcement, life insurance companies, school districts, and state agencies, are not required to comply with HIPAA protections. Importantly, the HIPAA Privacy Rule does not place restrictions on using or disclosing PHI of de-identified data. For defining whether data is sufficiently de-identified for disclosure, the U.S. Department of Health and Human Services (HHS) has defined two acceptable methods [61]. One is expert determination § 164.514 (b)(1), which requires applying statistical or scientific principles and is used by the Department of Veterans Affairs (VA) [3] when handling genomic data. The second is the "Safe Harbor Method" § 164.514 (b)(2), in which data is considered deidentified if 18 types of identifiers are removed and there is "no actual knowledge that the residual information can identify an individual." However, the second method is inappropriate for genomic data as genomic data is of extremely high dimensionality and even small fragments can re-identify individuals with a high degree of probability with current technology and publicly available information [3].

The Common Rule [62], officially known as the Federal Policy for the Protection of Human Subjects, establishes a standard of ethics for government-funded human subjects research. A 2017 update required all federally funded human subjects research to obtain meaningful informed consent. Study participants must be informed of how their genomic information will be used, who may access it, and what are the potential risks associated with release of PHI [3]. Human subjects participating in clinical studies funded by the NIH are automatically granted a Certificate of Confidentiality that safeguards their PHI. The Certificate of Confidentiality compels investigators and institutions to withhold PHI from civil or criminal proceedings. These certificates aim to promote clinical study participation by assuring a certain level of privacy. Non-federally funded studies, however, are not obligated to provide Certificates of Confidentiality.

4.3.2. Industry Resources

Future of Privacy Forum (FPF) Best Practices [56] is an industry supported voluntary set of principles targeted at the DTC genomic testing market and uses a Fair Information Practice Principles (FIPPs) based framework. It explicitly recognizes that genomic data absent other identifiers can usually be re-identified and thus continues to need strong protection provided by technical or contractual controls. The framework lists several important privacy issues and uses the FIPPs principles to address them at a high-level. It lacks specifics on cybersecurity and risk management of genomic data and is limited in focus to DTC genetic testing companies.

4.3.3. International Resources

Global Alliance for Genomics and Health: Data Privacy and Security Policy [57] is a document focused on the responsible data sharing of genomic and health-related data. It sets forth some general policies based on the four principles of (1) Respect Individuals, Families, and Communities, (2) Advance Research and Scientific Knowledge, (3) Promote Health, Wellbeing, and the Fair Distribution of Benefits and (4) Foster Trust, Integrity, and Reciprocity. It is an international guideline and does not explicitly consider U.S. privacy regulations.

4.4. Summary of Gaps in the Protection of Genomic Data

The NCCoE engaged with the public to identify common challenges, what is unique about genomic data, solutions that are available, and gaps faced by those who produce, store and process genomic data. Subject matter experts with experience in bioeconomy, cybersecurity, privacy, sequencing technologies, cloud hosting, and computation of genomic data were contacted. A five-and-a-half-hour virtual workshop was held on January 26, 2022, with nearly 500 stakeholders to understand the threats related to the privacy and security of genomic data and identify gaps in protection. Findings from that workshop were presented at the Healthcare Information and Management Systems Society (HIMSS) Global Health Conference & Exhibition on March 16, 2022, where feedback from attendees was solicited. An additional, second virtual workshop was held on May 18 and 19, 2022, where possible solutions were explored, and additional gaps were identified.

This section summarizes the gaps identified in guidance, technical solutions, and policy/regulations.

4.4.1. Guidance Gaps

Current cybersecurity and privacy risk management guidance does not address the specific and unique requirements of genomic data. This highlights the opportunity to address the following gaps:

- Publish voluntary guidance for protecting genomic data through the development of NIST Cybersecurity Framework and NIST Privacy Framework Profiles.

- Provide expertise to help life-sciences researchers leverage NIST guidance that supports FISMA and related requirements to improve the cybersecurity and privacy risk management of genetic data.
- Publish cybersecurity benchmarks or STIGs for the secure configuration of commercially available sequencers and the associated data analysis pipelines to address the current gap in visibility into how sequencer operating system and software settings have been configured for system hardening and the ability to assess these configurations after maintenance or updates.

4.4.2. Technical Solution Gaps

- Currently, sequencer manufacturers do not provide SBOMs for their devices. Therefore, security professionals have no visibility into potential vulnerabilities of their software and thus cannot adequately advise users of sequencers on how to address discovered software vulnerabilities through patching or other mitigation measures.
- Sequencers are typically connected to a network and the internet. This provides access to the manufacturer for updates and transfer of files to secure storage. There is no guidance on the network addresses and the corresponding network protocols that are required for sequencers to effectively operate. If this guidance existed, it would be possible to microsegment sequencers on the network, providing them with only the required resources for their proper functioning in keeping with ZTA cybersecurity principles. This would mitigate the possibilities of adversaries exploiting vulnerabilities in the sequencer's hardware or software for exfiltration of data as well as using the sequencers as entry points that allow lateral movement throughout the enterprise network.
- The general problem of data confinement, (i.e., authorized users and/or their software sharing unauthorized access to data), is a well-known unsolved problem in cybersecurity. Due to the privacy risks to subjects as well as the high value of many types of genomic data, the confinement problem is of relevance to genomic data. This problem is most commonly addressed by contractual controls that can be particularly complex when the controls are between multiple organizations. Further, contractual controls typically do not prevent unauthorized data sharing, but provide penalties if it is done. These penalties typically cannot redress the privacy loss of patients and data subjects.
- Most genetic data sharing and processing occurs in cloud environments, frequently leveraging containers (e.g., Docker or Pods). Many cybersecurity vulnerability scanners are not optimized for scanning containers, resulting in an inability to identify certain vulnerabilities and a high number of false positives.
- In the healthcare of a patient, the nature and size of genomic data presents a challenge with incorporating it into EHR workflows. Fast Healthcare Interoperability Resources (FHIR) as a genomic cybersecurity solution provides important security features, such as tracking provenance and consent as well as providing encryption for privacy. However, additional work is needed to scale for genomic data size files.

4.4.3. Policy/Regulatory Landscape

- While some forms of information are subject to export controls, U.S. laws do not treat genetic data as a national security asset, but primarily focus on privacy and intellectual property (IP) protection. Few restrictions prevent a U.S. company from selling genetic data to parties outside the U.S. Some of these gaps may be addressed through efforts to implement EO 14081.
- Data held by DTC genetic testing companies are not subject to HIPAA or privacy and security requirements that apply to health care providers, as consumers send samples directly to the companies without the involvement of a healthcare provider. Only recently have several states enacted consumer protection laws directed at DTC genetic testing companies. These laws provide consumers with additional rights such as requiring consent for data sharing or granting consumers the ability to access or delete their data [63][64][65].
- Because laws of some countries may prevent the aggregation of a global human genome representing all people groups [45][48], AI methods trained on genomic datasets may be biased with respect to U.S. citizens whose heritage includes portions of the under-represented people groups.
- Multiple peer-reviewed studies [6][7][8][28] have demonstrated that even small parts of genomic data can be re-identified with high probability of success. The NIH and the VA consider the re-identification risk of genomic data and have department-specific policies to prevent this occurrence [3]. Existing de-identification approaches, like the HIPAA safe harbor provision, do not fully circumvent re-identification risk because genomic data alone may be sufficient to re-identify an individual. Moreover, HIPAA only applies to covered entities and genomic information in the context of PHI.

5. Available Solutions to Address Current Needs

This section describes opportunities to address the gaps identified in Section 4 through technologies, processes, and guidance.

5.1. NIST Cybersecurity Framework Profile for Processing of Genomic Data

Genomic data is highly valuable to organizations in the bioeconomy. Organizations need to apply appropriate cybersecurity capabilities for the protection of the data. The NIST Cybersecurity Framework is a voluntary, comprehensive, and applicable framework that organizations can use as part of an overall risk management plan. However, the NIST Cybersecurity Framework is a general framework, and many industries have found it beneficial to have a Cybersecurity Framework Profile based on that industry's mission objectives.

An industry-specific Cybersecurity Framework Profile provides several benefits. A Profile can be used as a standardized approach for preparing a cybersecurity plan appropriate for the security of genomic data; a method to select controls and mitigations appropriate for the high value of genomic data; and can be used for gap analysis for organizations already storing and processing genomic data to examine their cybersecurity posture against and prioritize upgrades to that posture. Thus, a Cybersecurity Framework Profile specifically tailored to the mission objectives of those who handle genomic data would be a valuable addition for securing the bioeconomy. The NCCoE engaged stakeholders from the genomic community to create the Cybersecurity Framework Profile for genomic data and anticipates publishing the Profile in the summer 2023.

5.1.1. Use Case Description

Organizations that handle genomic data need to protect that data due to both its high value as well as the privacy risk to individuals if the data were exposed. The organization needs to select appropriate controls to reduce the risk to the confidentiality, integrity, and availability of the data to an acceptable level. An organization must develop an appropriate cybersecurity plan to accomplish the task in a cost-effective manner so that the mission objectives of the organization can be achieved. After implementing the plan, the organization needs to periodically assess its cybersecurity posture considering new technology and threats and to analyze their current posture versus an appropriate target posture to determine if there are gaps in its cybersecurity posture and prioritize the remediation of those gaps.

5.1.2. Solution Idea

There exist a number of Target Profiles for specific use cases for the NIST Cybersecurity Framework [66], however none address the bioeconomy, considering its challenges and mission objectives. A NIST Cybersecurity Framework Profile for genomic data would be particularly relevant because of the high value of genomic data and the important risks to the nation, economy and individuals' loss of that data can have. The Profile could highlight controls that address the unique aspects of genomic data and the most important considerations for organizations that process genomic data.

5.1.3. Expected Benefits

A Cybersecurity Framework Profile that is specific to genomic data would assist organizations who aggregate and process genomic data by highlighting gaps in their cybersecurity capabilities. It would also assist organizations in prioritizing and implementing additional capabilities or controls.

5.2. NIST Privacy Framework Profile for Processing Genomic Data

Human subjects who provide their genomic data for research expect that their privacy will be protected by all organizations involved in the research process. Human genomic data can reveal a great deal of personal information, such as physical traits, predisposition to certain health conditions, and biological relationships. Individual privacy may be impacted when an individual is identified, and other protected or sensitive information is uncovered or made available. For instance, identifying an individual in a genomics database for patients with a particular medical condition may impact the privacy of that individual by revealing sensitive health information. The genetic data itself can serve as an identifier and provide additional information about the contributor. Lastly, aggregate data in large genomic databases can lead to the identification of individuals or their biological relatives [8][68]. Deidentifying genomic data is impossible without destroying some or all of utility of that data [3].

In addition to a broad range of privacy considerations, human subjects provide their data under informed consent, which may further limit the processing of their data to specific uses. Genomic data collected with differing informed consents is often aggregated, but when this is done, care must be taken that all data is still used within the boundaries specified in the informed consent of each data subject. Organizations that process genomic data need to be able to effectively communicate internally and externally about managing these and other privacy risks.

The NIST Privacy Framework [52], a voluntary tool, helps organizations to prioritize the policies and technical capabilities they need to manage the privacy risks that may arise from data processing, including processing genomic data. Like the Cybersecurity Framework, the Privacy Framework can be applied to and tailored for the mission objectives for processing genomic data to manage risks to individuals as well as related risk that can arise to organizations when developing their products, systems, and services (e.g., reputational risk, loss of trust, financial risk).

5.2.1. Use Case Description

Organizations that process human genomic data need to protect individuals from privacy risk. These organizations also need to process their genomic data within the boundaries allowed by each subject's informed consent and usually must respect and manage multiple different informed consents. They must have a plan for effective technical controls and processes that will reduce the privacy risk to an acceptable level while still accomplishing the mission objective in a timely and cost-effective manner.

5.2.2. Solution Idea

There exist a number of Target Profiles for specific use cases for the NIST Cybersecurity Framework [66]; however, no NIST Privacy Framework Profiles have been developed. A NIST Privacy Framework Profile would be particularly relevant for genomic data because of the high sensitivity of genomic data and the unique reidentification risks associated with genomic data. The Profile could address the unique aspects of genomic data and highlight the most important considerations for aggregators and processors of genomic data.

5.2.3. Expected Benefits

A Privacy Framework Profile that is specific to genomic data would assist organizations who aggregate and process genomic data by highlighting gaps in capabilities intended to protect the privacy of individuals. It would also assist organizations in prioritizing implementation of additional privacy capabilities or controls.

5.3. Automatic Network Micro-Segmentation of Sequencers with MUD

The NCCoE recently developed model implementations of a solution to help secure internet of things (IoT) devices that may have applications for genomic sequencers. NIST released NIST SP 1800-15, Securing Small-Business and Home Internet of Things (IoT) Devices: Mitigating Network-Based Attacks Using Manufacturer Usage Description (MUD) in 2021 [68]. MUD restricts a managed device to communicate with only an "allowlist" of permitted network addresses (internet protocols and ports) specific to each type of device, consistent with ZTA cybersecurity principles.

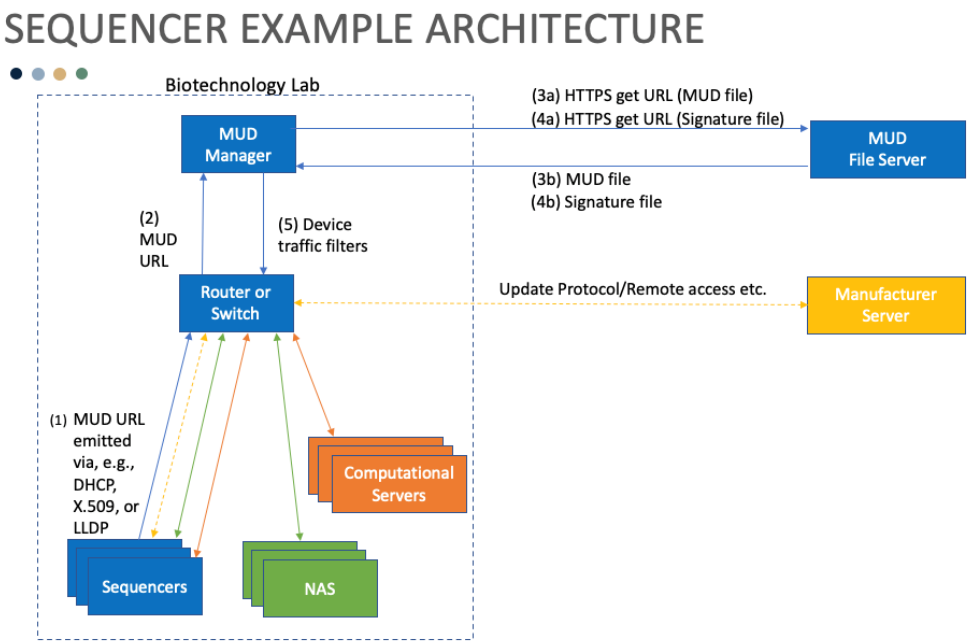
5.3.1. Use Case Description

Sequencers are expensive devices that are operated by custom software that provides only limited visibility into network connections and configurations. Sequencers must be connected to the internet and intra-network for manufacturer service, user access, and data storage, but the number of communication addresses and protocols they require for proper operation are limited. In these ways, sequencers can be compared to IoT devices [3]. Because of sequencers' connectivity, they may be an attractive target for data exfiltration, ransomware deployment, malware implantation, and lateral movement within the user's enterprise network.

5.3.2. Solution Idea

MUD could be applied to sequencers. Since sequencers have a small number of communication patterns, MUD enables a network to limit sequencer communication within the local network and externally only to those resources needed for proper functionality. There is a MUD architecture that implements MUD and contains a MUD Manager that provides micro-segmentation of the device based on the manufacturer-provided allowlist. This micro-segmentation provided by the MUD Manager greatly inhibits adversaries' ability to access a managed device; and if they do access a device, their ability to move laterally across the rest of

984 the network is severely restricted by the MUD Manager. Figure 3 illustrates a potential MUD
985 architecture for a sequencer [4].



986 **Fig. 3.** Notional MUD architecture for a sequencer

987 MUD can be applied in partnership with the manufacturer. Existing tools can help either a
988 manufacturer or a third party develop an allowlist implemented using MUD for a device that
989 controls a device's network traffic. A significant advantage of the MUD solution is that it can
990 also be implemented with legacy devices (whose software cannot be modified) to provide the
991 needed MUD information when connected to the network.

992 ### 5.3.3. Expected Benefits

993 A model solution with sequencers could facilitate MUD's adoption within the bioeconomy and
994 significantly improve security across a large range of stakeholders including commercial,
995 academic, and government organizations.

996 By increasing security across multiple stakeholders, MUD would reduce the likelihood for
997 ransomware attacking and preventing usage of these valuable assets, as well as possible
998 intellectual property loss or privacy loss from exfiltration of data.

999 ## 5.4. Security Guidelines for Data Analysis Pipelines

1000 Security guidelines (such as DISA STIGS or configuration baselines) manage and reduce
1001 cybersecurity risk by providing consensus-based and auditable configurations and security
1002 features for systems. They are a well-established method of providing transparency and
1003 alignment between user requirements and manufacturers' product offerings. These guidelines

may be used to harden data analysis pipelines and devices (for example, sequencers) to a trusted configuration baseline.

5.4.1. Use Case Description

Data analysis pipelines interact with a combination of open-source, off-the-shelf, and custom software including operating systems. Changes to software or configurations must be carefully validated and verified to ensure that they do not affect the pipelines' operation or introduce cybersecurity or privacy risks. Data analysis pipelines, including sequencers, typically prioritize ease of use and accuracy of results over minimizing cybersecurity and privacy risks. As such, it has been demonstrated that these pipelines and sequencers can be susceptible to adversarial threats [69].

5.4.2. Solution Idea

Security guidelines for data analysis pipelines (including sequencers) would provide best practice cybersecurity hardening guidance. These customized guidelines could be based off existing STIGs [70] or Center for Internet Security (CIS) Benchmarks [71] and tailored for the unique requirements of sequencers.

These security guidelines might be modeled after NISTIR 8259 [72] for IoT devices that enumerates the capabilities, features, and functionalities that manufacturers of sequencers need to provide so that users can mitigate their cybersecurity risks. Security guidelines could include a requirement for the SBOM detailing all software installed on the sequencer to assist in identifying and mitigating cybersecurity vulnerabilities [58].

5.4.3. Expected Benefits

By establishing these guidelines and demonstrating feasibility, users would be able to include security standards in their purchasing requirements and sequencer manufacturers would have clear guidance on how to achieve the security standards. It would also enable assessing delivered products, enabling a device to be more quickly put into use both at initial delivery and after upgrade or service. This solution enables increased commerce by aligning user cybersecurity needs with manufacturers' cybersecurity offerings. The reduced friction from eliminating differing user expectations and manufacturer offerings would result in fewer purchasing delays and more competition between manufacturers on relevant user cybersecurity needs. It would also allow users to better manage their security risks in accordance with Federal information security requirements.

5.5. Demonstration Project for Genomic Data Risk Management

Because of the high value of genomic data, organizations in the bioeconomy and researchers using genomic data need to apply appropriate risk management. The NIST Risk Management Framework suite of guidance documents are comprehensive and can be applied to genomic data.

A demonstration project providing an example of how they can be used along with the resources required for the effort would be beneficial.

There does not exist a roadmap or demonstration project specifically tailored to the unique needs of aggregating and processing genomic data. These needs include consent management and reidentification risk. Consent management can be particularly complex in the large aggregations of subject data that are typical in oncology and precision medicine. The genomic data is typically pooled from many studies that may be collected with differing informed consent. The informed consent of each subject needs to be tracked and each analysis needs to be cleared for specific informed consents. The reidentification risk from genomic data stems from the characteristic that deidentification cannot be effectively done without sacrificing the utility of the data. If reidentification is successful, the subject data can be used for kinship, phenotype prediction, health information, and other possible future applications that may cause harm to the subject or the subject's kin.

5.5.1. Use Case Description

The target stakeholder would be a small organization, such as a university or start-up, that wants to aggregate and process genomic data. A start-up has important intellectual property that needs to be protected. A university may want to utilize NIH genomic data and must implement NIH cybersecurity and privacy requirements. Both need a solution to manage the risk of aggregating and processing genomic data that can be done in a timely and cost-effective manner.

5.5.2. Solution Idea

The demonstration project will identify controls, aligned to NIST SP 800-53, that are in place in a vendor-proposed solution that align to the unique requirements of handling genomic data. Two solutions exist that map their security implementation to NIST SP 800-53, Terra.bio and DNAnexus. However, organizations must understand the unique requirements for handling genomic data. This does not alleviate the responsibility of the organization to implement appropriate risk management but can streamline the effort by using existing resources.

5.5.3. Expected Benefits

This demonstration project can identify gaps in vendor-proposed solutions aligned to genomic data handling cybersecurity and privacy requirements. The project could document savings in time, person hours, required expertise and documentation, to help an organization properly plan and budget their risk management effort. The guidance documentation produced from this effort would be specific to genomic data and could assist users by reducing mistakes due to missed requirements or misunderstanding of the requirements.

5.6. Demonstration Project for Analysis of Genomic Data Using Privacy Preserving and Enhancing Technologies

There are several promising technologies that may assist in addressing data subject privacy [3][74][75] and the breach of data confinement in genomic data sharing and analysis. Sharing high-value data typically requires a high degree of trust between data sharing organizations with significant contractual agreements. The process of negotiating and agreeing to the contractual arrangements and necessary controls can greatly slow and, in many cases, prohibit the sharing of data.

Solutions that have been used or proposed for genomic datasets for addressing subject data privacy and preventing the breach of data confinement include:

- A genomic analysis platform brings researchers to a secure location, requiring the use of plaintext and blocking all egress. This is not typically practical and severely limits research because all researchers must come to the secure location.
- A "Secret Store" requires the use of predefined executables with no direct visibility to the data. However, this severely limits analysis options (only previously defined, validated, and verified executables can be used) available to researchers [4].
- All data is stored so that analysis is done in a cloud environment created by a trusted authority, and all activity in that environment is monitored. This is not a complete solution, as researchers can provide visibility and/or their authorization to others. Additionally, this removes the advantage of remote access service which limits authorized users to only what they are authorized to do with the data, as currently available cloud solutions use identity-based authorization.
- Federated Machine Learning (FML) [76], where collaborative learning is done among individual local nodes and shared with a centralized processor to produce an aggregate model, has been shown to be a promising method in precision medicine. It provides the capability to train models on competing or otherwise separate entities to produce an average model that captures the patterns present in the local nodes. However, unless combined with other privacy-enhancing technologies and controls, it can be vulnerable to data reconstruction attacks.[2]
- Differential privacy can provide privacy guarantees and be combined with FML [77] or used to produce synthetic data [78] for machine learning. Differential private synthetic data adds noise to datasets so that an individual's data cannot be distinguished within the dataset and allows the data to be analyzed and shared. Additionally, care must be taken when sharing this data, particularly if combining it with other data sets, as there is not yet sufficient research to determine whether combining it with other datasets maintains the privacy guarantee provided by differential privacy. It also allows for periodic updates as

[2] Privacy enhancing technologies or PETs include tools like homomorphic encryption, secure multi-party computation, and differential privacy. This is an evolving area and while some tools show promise, they are not a complete solution on their own and must be deployed with other controls. NIST is one of the organizers of the U.K.-U.S.PETs Prize Challenges which has a goal of "Accelerating the adoption and development of PETs." Additional information and results are available at: https://petsprizechallenges.com/.

the data evolves because differential private synthetic data can cope with the progress of the underlying dataset seamlessly. Researchers, however, have expressed concerns with the accuracy of the synthetic data as there is a tradeoff depending on where the privacy parameter is set. There are significant technology improvements in development to increase the accuracy of the synthetic data and some analyses are not as sensitive to the accuracy of the synthetic data.

- Much progress has been made in privacy enhancing cryptography (PEC) [79] which addresses both data confinement and subject data privacy. Fully homomorphic encryption (FHE) allows computation on encrypted data and progress has been rapid at addressing its principal drawback of increased computational complexity. Another promising PEC technique is secure multi-party computation (SMPC) where analysis is performed over the data sets of several parties without revealing their input. Finally, there may be a place for zero-knowledge proofs (ZKP) [80][81] for validating results of genomic analysis without revealing the solution, which may fully preserve the privacy of the data underpinning the results.

5.6.1. Use Case Description

Large genomic datasets are required in oncology and precision medicine. Multiple dataset aggregation is often required. This typically requires the negotiation of multiple contracts between all the sharing organizations and the complexity of the contracts, since the datasets are extremely confidential, can be prohibitive. If data could be shared without a concern about exposure or exfiltration, the complexity of such contracts would be greatly simplified, and risk greatly reduced.

5.6.2. Solution Idea

Within the field of genomics, a solution that can virtually eliminate the risk of confidentiality or integrity loss of sharing genomic data between organizations and solve the confinement problem is federated multi-party homomorphic encryption [82]. This is a technique that allows for computation on encrypted data aggregated over multiple datasets. Exfiltration of the raw data is prevented as the authorized user is only able to obtain results from the computation which involves multiple datasets and authorized users cannot access the raw data in plaintext.

At this stage in technology development, federated homomorphic encryption is not a general solution. It has been proven useful for several important problems in the analysis of genomic data, particularly in oncology and precision medicine research. In oncology, it has been shown to allow survival analysis that enables the effectiveness of different treatment options based on the genomic information of the patient and/or tumor to be compared [82]. In precision medicine, it allows genome-wide association studies (GWAS) [82] as well as machine learning (ML) training and testing. The homomorphic encryption solution is being used in Switzerland. A demonstration project in the United States working with The Broad Institute (United States) and/or Tune Insight (Switzerland) would be helpful for the technique to gain wider acceptance in the U.S.

5.6.3. Expected Benefits

This solution can enable more rapid progress in oncology treatment and precision medicine by allowing collaboration between organizations without complex contract negotiations or sharing agreements because the risk of exfiltration of data or privacy violation is virtually eliminated. Additionally, patient and research subjects' privacy would be controlled by technical means, which can be less prone to failure and data breaches compared to contractual controls.

6. Areas for Further Research

There are several problems where more research is needed that may be fruitful areas for NIST or NCCoE sponsored research work to address:

- FedRAMP assessors use vulnerability scanners to assess the security posture of systems. Container (e.g., Docker or Pods) cybersecurity vulnerability assessment scanners are plagued with a high number of false positives. Research on how to improve the ability of vulnerability scanners to identify security issues in containers would benefit security professionals who maintain systems processing genomic data and optimize the use of their cybersecurity resources.
- Research on how to securely integrate whole genomic data with a patient's EHR while allowing for privacy and interoperability is needed. The FHIR standard provides a framework, but a genomic specific implementation is needed along with a demonstration project.
- While federated multi-party homomorphic encryption can solve the confinement problem for a class of analyses, more work is needed as there are analysis methods not addressed by this solution. For the last few years, NIH has funded iDASH (integrating data for analysis, anonymization, and sharing) [74] and held secure genome analysis competition workshops to study specific cybersecurity mechanisms for genomic data systems. The key focus of this annual workshop is to advance cybersecurity technologies that are essential for genomic and biomedical system security and privacy. While such technical advancement is crucial for the security and privacy of genomic data sharing, more holistic and systematic security solutions that may use those technologies are also important for addressing systemic issues found in genomic information systems.

7. Conclusion

This document describes the challenges and concerns associated with handling genomic data, the current state of relevant cybersecurity and privacy risk management practices, gaps in implementing genomic data protections, and potential solutions along with areas for further research. Because of the value of genomic data, the associated challenges and concerns may affect national security, the U.S. economy, intellectual property, individual privacy, and under-protected populations. Existing cybersecurity and privacy risk management practices such as the NIST RMF, Cybersecurity Framework, and Privacy Framework must be tailored to effectively implement appropriate genomic data protections. Additionally, gaps persist in current policy, legislation, technology, and guidance for protecting genomic data.

Solutions identified to address these challenges and gaps include:

- A NIST Cybersecurity Framework Profile for handling of genomic data could help organizations identify gaps and prioritize investments in cybersecurity capabilities and controls.
- A NIST Privacy Framework Profile for handling of genomic data could provide clarification on how to manage the privacy risks inherent in the aggregation, storage, and processing of human genomic data.
- The Manufacturer Usage Description specification could improve sequencer security and reduce the likelihood of ransomware attacks as well as intellectual property loss or privacy loss from exfiltration of data.
- Security guidelines or benchmarks for sequencers could provide best practice cybersecurity hardening guidance, require SBOMs to improve supply chain security, and improve cyber resiliency against future threats.
- A demonstration project highlighting the benefits of using RMF guidance for protecting genomic data could illustrate how organizations can leverage appropriate secured cloud-based solutions to reduce the time, person hours, required expertise, and documentation required to implement effective cybersecurity and privacy practices.
- A federated homomorphic encryption demonstration project for analysis of genomic data in precision medicine or oncology could illustrate how these solutions reduce the risk of confidentiality or integrity loss when sharing genomic data between organizations and help address the confinement problem.

Future research areas may include methods for securely integrating whole genomic data with a patient's EHR while allowing for privacy and interoperability; improving the precision of vulnerability scanners for software containers; and technical solutions to solve the containment problem in genomic data for analysis methods not currently addressed by federated multi-party homomorphic encryption.

References

[1] Executive Office of the President, United States, "Executive order 14801: Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy," *Federal Register 56849-56860,* vol. 87, no. 178, 15 September 2022.

[2] National Institute of Standards and Technology (NIST), "NIST Privacy Framework, Version 1.0," 16 January 2020. [Online]. Available: https://doi.org/10.6028/NIST.CSWP.01162020.

[3] National Cybersecurity Center of Excellence, "NCCoE Virtual Workshop on the Cybersecurity of Genomic Data," 26 January 2022. [Online]. Available: https://www.nccoe.nist.gov/get-involved/attend-events/nccoe-virtual-workshop-cybersecurity-genomic-data/post-workshop-materials.

[4] National Cybersecurity Center of Excellence, "NCCoE Virtual Workshop on Exploring Solutions for the Cybersecurity of Genomic Data," 18 May 2022. [Online]. Available: https://www.nccoe.nist.gov/get-involved/attend-events/nccoe-virtual-workshop-exploring-solutions-cybersecurity-genomic-data.

[5] M. Gymrek and et.al., "Identifying Personal Genomes by Surname Inference," *Science,* vol. 339, pp. 321-324, January 2013.

[6] M. Naveed and et.al., "Privacy in the Genomic Era," *ACM Computer Survey,* vol. 48, no. 1, pp. 1-49, September 2015.

[7] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews,* vol. 15, pp. 409-421, June 2014.

[8] Y. Erhlich and et.al., "Identity Inference of Genomic Data Using Long-Range Familial Searches," *Science,* vol. 362, pp. 690-694, 2018.

[9] C. A. Hutchison, "DNA sequencing: bench to bedside and beyond," 12 September 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2094077/. [Accessed 15 December 2021].

[10] National Institutes of Health (NIH), "DNA Sequencing Costs: Data," 1 November 2021. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. [Accessed 15 December 2021].

[11] NIH National Center for Biotechnology Information (NCBI), "Genome Reference Consortium," [Online]. Available: https://www.ncbi.nlm.nih.gov/grc/human. [Accessed 18 July 2022].

[12] J. Wagner and et.al., "Benchmarking challenging small variants with linked and long reads," *Cell Genomics,* vol. 11, no. 2, p. 100128, 2022.

[13] J. Chen and et.al., "Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers.," *Scientific Reports,* vol. 9, no. 1, pp. 1-13, 2019.

[14] Z. Chen and et.al., "Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency.," *Scientific Reports,* vol. 10, no. 1, pp. 1-9, 2020.

[15] Market Research Guru, "Global And United States Direct To Consumer Dtc Genetic Testing Market - Industry Reports," 20 May 2022 [Online]. Available: https://www.marketresearchguru.com/global-and-united-states-direct-to-consumer-dtc-genetic-testing-market-20993717. [Accessed 20 July 2022].

[16] NIH National Center for Biotechnology Information (NCBI), "Summary Statistics of dbGaP Data," [Online]. Available: https://www.ncbi.nlm.nih.gov/projects/gap/summaries/cgi-bin/molecularDataPieSummary.cgi. [Accessed 20 July 2022].

[17] M. O'Brien, "Who Has the Largest DNA Database? (2022)," 17 July 2021. [Online]. Available: https://www.dataminingdna.com/who-has-the-largest-dna-database/. [Accessed 15 July 2022].

[18] T. Callaghan, "Responsible genetic genealogy," *Science,* vol. 366, no. 6462, p. 155, 2019.

[19] R. Kain and et.al., "Database shares that transform research subjects into partners," *Nature biotechnology,* vol. 37, no. 10, pp. 1112-5, 2019.

[20] K. Kozminski, "Biosecurity in the age of Big Data: a conversation with the FBI," *Molecular Biology of the Cell,* vol. 26, pp. 3894-3897, 2015.

[21] National Counterintelligence and Security Center (NCSC), "China's Collection of Genomic and Other Healthcare Data From America: Risks to Privacy and U.S. Economic and National Security," pp. 1-5, February 2021.

[22] A. Cohen, "MGH data breach exposes 10,000 patients," *Boston Herald,* 22 August 2019.

[23] N. Grant, "DNA testing service Vitagene exposed thousands of customer records online for years," *Los Angeles Times,* 9 July 2019.

[24] MyHeritage, "Security alert: malicious phishing attempt detected, possibly connected to GEDmatch breach," 21 July 2020. [Online]. Available: https://blog.myheritage.com/2020/07/security-alert-malicious-phishing-attempt-detected-possibly-connected-to-gedmatch-breach/. [Accessed 16 November 2021].

[25] "Devious 'Tardigrade' Malware Hits Biomanufacturing Facilities," 22 November 2021. [Online]. Available: https://www.wired.com/story/tardigrade-malware-biomanufacturing/. [Accessed 11 July 2022].

[26] M. Foley, "Genetically Engineered Bioweapons: A New Breed of Weapons for Modern Warfare," *Dartmouth Undergraduate Journal of Science,* vol. Winter, 2013.

[27] J. Peccoud and et.al., "Cyberbiosecurity:From Naive Trust to Risk Awareness," *Trends in Biotechnology,* vol. 36, no. 1, pp. 4-7, January 2018.

[28] A. Schwab and et.al., "Genomic Privacy," *Clinical Chemistry,* vol. 64, no. 12, pp. 1696-1703, 2018.

[29] M. Gianfrancesco and et.al., "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine,* vol. 11, no. 178, pp. 1544-1547, 2018.

[30] Y. Joly and et.al., "Establishing the International Genetic Discrimination Observatory," *Nature Genetics,* vol. 52, pp. 466-468, 2020.

[31] R. Parikh and et.al., "Addressing Bias in Artificial Intelligence in Health Care," *JAMA,* vol. 24, no. 322, pp. 2377-2378, 2019.

[32] P. Grother and et.al., "NISTIR 8280 Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects," December 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf. [Accessed 28 July 2022].

[33] CBS News, "The clones of polo," [Online]. Available: https://www.cbsnews.com/news/the-clones-of-polo. [Accessed 20 July 2022].

[34] National Institute of Health/National Human Genome Research Institute, "Eugenics and Scientific Racism," 18 May 2022. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Eugenics-and-Scientific-Racism. [Accessed 19 July 2022].

[35] National Instititute of Standards and Technology (NIST), "SP 800-37 Rev. 2 Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy," December 2018. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-37/rev-2/final.

[36] Executive Office of the President, United States, "Improving the Nation's Cybersecurity," 12 May 2021. [Online]. Available: https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity.

[37] Global Alliance for Genomics and Health (GA4GH), "Framework for Responsible Sharing of Genomic and Health-Related Data," Global Alliance for Genomics and Health, 3 September 2019. [Online]. Available: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/. [Accessed 15 November 2021].

[38] U.S. Office of the Federal Register, "Public Law 113-282 - Federal Information Security Modernization Act of 2014," U.S. Government Publishing Office, 2014.

[39] "FedRAMP," U.S. General Services Administration, [Online]. Available: FedRAMP.gov. [Accessed 2023 01 18].

[40] National Institute of Standards and Technology (NIST), "NIST SP 800-161r1," May 2022. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1.pdf.

[41] Scott Rose and et.al., "NIST SP 800-27," August 2020. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-207/final.

[42] "CISA Zero Trust Maturity Model," Cybersecurity & Infrastructure Security Agency, 2022. [Online]. Available: https://www.cisa.gov/zero-trust-maturity-model.

[43] OneTrust, "The Definitive Guide to Schrems II," 25 March 2022. [Online]. Available: https://www.dataguidance.com/resource/definitive-guide-schrems-ii. [Accessed 28 July 2022].

[44] E. Rafaelof and et.al., "Translation: China's 'Data Security Law (Draft)'," 2 June 2020. [Online]. Available: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinas-data-security-law-draft/. [Accessed 2020 July 2022].

[45] N. Dhavate and M. Ramakant, "A look at proposed changes to India's (Personal) Data Protection Bill," 16 December 2021. [Online]. Available: https://iapp.org/news/a/a-look-at-proposed-changes-to-indias-personal-data-protection-bill/. [Accessed 27 July 2022].

[46] India Ministry of Electronics and Information Technology, "THE PERSONAL DATA PROTECTION BILL, Bill No. 373 of 2019," 2019. [Online]. Available: http://164.100/47/4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf. [Accessed 1 Aug 2022].

[47] R. Creemers and et.al., "Translation: Cybersecurity Law of the People's Republic of China (Effective June 1, 2017)," New America, 29 June 2018. [Online]. Available: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-cybersecurity-law-peoples-republic-china/. [Accessed 1 August 2022].

[48] D. Cyranoski, "China announces hefty fines for unauthorized collection of DNA," *Nature News,* 14 June 2019.

[49] InCountry Staff, "Russian Data Protection Laws: Essential Guide on Compliance Requirements in Russia," InCountry, 19 March 2021. [Online]. Available: https://incountry.com/blog/russian-data-protection-laws-essential-guide-on-compliance-requirements-in-russia/. [Accessed 1 August 2022].

[50] OneTrust, "Russia - Data Protection Overview," OneTrust, April 2022. [Online]. Available: https://www.dataguidance.com/notes/russia-data-protection-overview. [Accessed 1 August 2022].

[51] Global Alliance for Genomics and Heath (GA4GH), “Data Security Infrastructure Policy,” 21 October 2019. [Online]. Available: https://github.com/ga4gh/data-security/blob/master/DSIP/DSIP_v4.0.md.

[52] National Institute of Standards and Technology (NIST), "Cybersecurity Framework Version 1.1," 16 April 2018. [Online]. Available: https://doi.org/10.6028/NIST.CSWP.04162018. [Accessed March 2022].

[53] National Institute of Standards and Technology, "NIST Special Publication 800-207 Zero Trust Architecture," 11 December 2020. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-207/final.

[54] "Security Technical Implementation Guides (STIGs)," DoD Cyber Exchange Public, [Online]. Available: https://public.cyber.mil/stigs/. [Accessed 18 January 2023].

[55] "bioisac," 2022. [Online]. Available: https://www.isac.bio/. [Accessed 18 January 2023].

[56] Future of Privacy Forum, "Privacy Best Practices for Consumer Genetic Testing Services," 31 July 2018. [Online]. Available: https://fpf.org/wp-content/uploads/2018/07/Privacy-Best-Practices-for-Consumer-Genetic-Testing-Services-FINAL.pdf. [Accessed 2 December 2021].

[57] Global Alliance for Genomics and Health (GA4GH), "Privacy and Security Policy," 26 May 2015. [Online]. Available: https://www.ga4gh.org/wp-content/uploads/Privacy-and-Security-Policy.pdf.

[58] Cybersecurity & Infrastructure Security Agency (CISA), "Software Bill of Materials," [Online]. Available: https://www.cisa.gov/sbom. [Accessed 10 June 2022].

[59] 110th U.S. Congress, "Public Law 110-233," 21 May 2008. [Online]. Available: https://www.govinfo.gov/content/pkg/PLAW-110publ233/pdf/PLAW-110publ233.pdf. [Accessed 2 December 2021].

[60] 104th U.S. Congress, "Health Insurance Portability and Accountability Act of 1996," 20 August 1996. [Online]. Available: https://aspe.hhs.gov/reports/health-insurance-portability-accountability-act-1996. [Accessed 2 December 2021].

[61] Department of Health and Human Services (HHS), "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard. [Accessed 10 June 2022].

[62] Department of Health and Human Services (HHS), "Federal Policy for the Protection of Human Subjects ('Common Rule')," 13 December 2022. [Online]. Available: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html. [Accessed 18 January 2023].

[63] Arizona State Legislature, "44-8002. Direct-to-consumer genetic testing company requirements; prohibition," 29 September 2021. [Online]. Available: https://www.azleg.gov/viewdocument/?docName=https://www.azleg.gov/ars/44/08002.htm. [Accessed 18 January 2023].

[64] California State Legislature, “Senate Bill No. 41, Chapter 596,” 7 October 2021. [Online]. Available: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220SB41. [Accessed 18 January 2023].

[65] Utah Legislative General Counsel, "Genetic Information Privacy Act," 22 February 2021. [Online]. Available: https://le.utah.gov/~2021/bills/sbillint/SB0227.pdf. [Accessed 18 January 2023].

[66] National Institute of Standards and Technology (NIST), "Cybersecurity Framework - Examples of Framework Profiles," NIST, 5 July 2022. [Online]. Available: https://www.nist.gov/cyberframework/examples-framework-profiles. [Accessed 7 July 2022].

[67] N. Homer and et.al., "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays," *PLoS Genetics,* vol. 4, no. 8, p. e1000167, 2007.

[68] National Institute of Standards and Technology (NIST), "NIST Special Publication 1800-15 Securing Small-Business and Home Internet of Things (IoT) Devices: Mitigating Network-Based Attacks Using Manufacturer Usage Description (MUD)," May 2021. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-15.pdf.

[69] C. Cimpanu, "Mysterious Iranian group is hacking into DNA sequencers," 14 June 2019. [Online]. Available: https://www.zdnet.com/article/mysterious-iranian-group-is-hacking-into-dna-sequencers/. [Accessed 13 June 2022].

[70] DoD Cyber Exchange Public, "STIGs Document Library," [Online]. Available: https://public.cyber.mil/stigs/downloads/. [Accessed 10 June 2022].

[71] Center for Internet Security, "CIS Benchmarks," [Online]. Available: https://www.cisecurity.org/cis-benchmarks/. [Accessed 10 June 2022].

[72] National Institute of Standards and Technology (NIST), "Foundational Cybersecurity Activities for IoT Device Manufacturers," May 2020. [Online]. Available: https://csrc.nist.gov/publications/detail/nistir/8259/final.

[73] National Institute of Standards and Technology (NIST), "Cyber-Physical Systems/Internet of Things Testbed," 7 April 2022. [Online]. Available: https://www.nist.gov/programs-projects/cyber-physical-systemsinternet-things-testbed. [Accessed 15 July 2022].

[74] "iDASH Privacy & Security Workshop," [Online]. Available: http://www.humangenomeprivacy.org. [Accessed 19 November 2021].

[75] X. Shi and X. Wu, "An overview of human genetic privacy," *Annals of the New York Academy of Sciences,* vol. 1387, no. 1, pp. 61-72, 2017.

[76] N. Rieke and et.al., "The future of digital health with federated learning," *npj Digital Medicine,* vol. 3, no. 119, 2020.

[77] M. A. Aziz and et.al., "Generalized Genomic Data Sharing for Differentially Private Federated Learning," *Journal of Biomedical Informatics,* no. https://doi.org/10.1016/j.jbi.2022.104113, 2022.

[78] J. Near and D. Darais, "Differentially Private Synthetic Data," 3 May 2021. [Online]. Available: https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data. [Accessed 28 July 2022].

[79] National Institute of Standards and Technology (NIST), "Privacy-Enhancing Cryptography PEC," 22 July 2022. [Online]. Available: https://csrc.nist.gov/projects/pec. [Accessed 28 July 2022].

[80] S. Goldwasser and et.al., "The Knowledge Complexity of Interactive Proof Systems," *Siam Journal of Computing,* vol. 18, no. 1, pp. 186-208, 1989.

[81] L. Lesavre and et.al., "Blockchain Networks: Token Design and Managment Overview," National Institutue of Standards and Technology, NISTIR 8301, 2021.

[82] D. Froelicher and et.al., "Truly privacy-preserving federated analytics for precision 1441 medicine with multiparty homomorphic encryption.," Nature Communications, vol. 12, no. 1442 5910, 2021.

Appendix A. List of Symbols, Abbreviations, and Acronyms

The following acronyms are used in this publication.

AI
Artificial Intelligence

BIO-ISAC
Bioeconomy Information Sharing and Analysis Center

CIS
Center for Internet Security

CISA
Cybersecurity and Infrastructure Security Agency

CRISPR
Clustered Regularly Interspaced Short Palindromic Repeats

CRISPR-Cas
CRISPR-Associated Protein

DbGaP
Database of Genotypes and Phenotypes

DISA
Defense Information Systems Agency

DNA
Deoxyribonucleic acid

DTC
Direct-To-Consumer

EEA
European Economic Area

EHR
Electric Health Record

EO
Executive Order

FBI
Federal Bureau Investigation

FedRAMP
Federal Risk and Authorization Management Program

FHE
Fully Homomorphic Encryption

FHIR
Fast Healthcare Interoperability Resources

FIPPs
Fair Information Practice Principles

FISMA 2014
Federal Information Security Modernization Act (2002)

FISMA 2002
Federal Information Security Management Act (2002)

FPF
Future of Privacy Forum

GA4GH
Global Alliance for Genomic Health

GINA
Genetic Information Nondiscrimination Act (2008)

GWAS
Genome-Wide Association Studies

HIMSS
Healthcare Information and Management Systems Society

HIPAA
Health Insurance Portability and Accountability Act (1996; amended 2013)

iDASH
Integrating Data for Analysis, Anonymization, and Sharing

IoT
Internet of Things

IP
Intellectual Property

MUD
Manufacturer Usage Description

NCBC
National Centers for Biomedical Computing

NCBI
National Center for Biotechnology Information

NCCoE
NIST National Cybersecurity Center of Excellence

NGS
Next-Generation Sequencing

NISTIR
NIST Internal Report

OMB
Office of Management and Budget

PEC
Privacy Enhancing Cryptography

PHI
Protected Health Information

PMO
Program Management Office

RMF
Risk Management Framework

RNA
Ribonucleic Acid

SBOM
Software Bill of Materials

SMPC
Secure Multi-Party Computation

SRA
Sequence Read Archive

STIGS
Security Technical Implementation Guides

The Common Rule
Federal Policy for the Protection of Human Subjects

VA
Department of Veterans Affairs

ZKP
Zero-Knowledge Proof

ZTA
Zero-Trust Architecture