Genetic Data Categorization

Data Risk Evaluation for Safe, Secure Use

bioisac
Creating Collaborative Threat Intelligence for the Bioeconomy

Background

Genetic data is being generated at an ever-increasing rate as the cost of its generation continues to drop exponentially. Genetic data is created from the sequencing of DNA of organisms and may even arise from synthetic organisms engineered by scientists. Genetic data can be of high value, and in some instances, even present unique security risks. For example, the genetic data from a full genome (genomic data) of a dangerous pathogen can be used to recreate that organism. On the other hand, not all genetic data is sensitive, and some genetic data, such as that from basic research, is of low value to hackers and has fewer security concerns.

As we work to secure the genetic data infrastructure, we must be aware of the types of genetic data that will be generated and stored and consider the risks associated with the data, now and in the future. Laboratory cybersecurity is notoriously difficult, and in some instances extreme security measures are required to protect against persistent and talented adversaries. Data categorization is the first step in inventorying and understanding assets during security risk assessments and is crucial in determining the appropriate levels of security needed to handle various types of genetic data.

The need to consider risks that could arise from the future application of this data is necessary because of the immutability of DNA and rapidly evolving threat landscape related to the exploitation of this information. The DNA of an organism remains for its lifetime, which, for genetic data like that from humans, remains relevant for decades. Advances in technology can introduce exploits, vulnerabilities, and/or privacy violations. One such example of this advancement is the exponential rate of Artificial Intelligence (AI) growth and the democratization of biotechnology tools. Biotechnology tools, such as those used for genome modification, can be combined with AI advancements to lower the threshold by which genetic data can be used to cause harm. Genetic data today may appear harmless, but with the introduction of new forms of technology, it may become available for misuse. When genetic data is generated, handled and shared, users must consider how it may become dual-use, used to do good as well as harm.

The goal of this document is to categorize genomic data based on characteristics or attributes. Data categorization is a common approach to managing information within an organization for security and other purposes. By organizing data according to criticality, potential impacts, and/or potential risks posed by handling the data, one can assign specific security controls to different categorizations of data. As concerns increase with more and more sensitive forms of genetic data or with the known exploitation(s), recommended security controls can be applied to protect assets in an adaptive way.

Data Categorization

The first step in this effort is to define the categories. There are many ways to approach data categorization. For example, data can be categorized as sensitive to an organization due to its intrinsic value in the case where there would be significant material loss if it were to be leaked or destroyed. Data can also be categorized by its type and content, for example if it is known to contain identifying information or known to reveal private medical information. Naturally, such categorization should and in some cases does drive regulatory oversight. Finally, categorization based on risk is also possible, and one that allows more dynamic blending of factors. Risk for the purpose of this document is a combination of a threat (a person or event that can cause harm), a vulnerability (a weakness that can be exploited by the threat), and a consequence (the damage caused if a threat exploits a vulnerability).

For simplicity and clarity, we have opted for three categories in this document: 1) Low Risk (green), 2) Medium Risk (amber), and 3) High Risk (red). These three categories mirror the traffic light protocol used in cybersecurity, allowing for an easily recognizable color scheme, and were determined by reflecting on the sensitivity and value of genetic data and corresponding risks when handling said genetic data. The categorization offered can be adjusted based on data attributes, intended use, and additional nuance or needed modifications to fit the needs of a unique context or use case.

Low Risk (green) - This category has the lowest risk of misuse. General cybersecurity hygiene is recommended. For example, basic user authentication (e.g., username and password) should be required to handle this data when it is not public. If possible, enable multi-factor authentication to avoid password re-use attacks. Determine and implement the lowest level of security controls.
Medium Risk (amber) - This category has intermediate risk to both generators and consumers of this data. Low Risk requirements and controls are inherited and implemented, and the Medium Risk data access should require multi-factor authentication for digital access and a key card for physical access, especially if the risk of the data can be increased through aggregation.
High Risk (red) - This category represents the highest tier of risk where a single or small number of records or small amounts of data represents harm to the public, generators of the data, or the downstream consumer(s). The High Risk category should include the highest available security measures, such as restricting connectivity or keeping data offline (“air-gapped”). Medium Risk and Low Risk requirements are inherited and associated controls are implemented.

Security Requirements

Once the categories and characteristics are defined, security requirements can be assigned to each category. Security requirements are a set of rules and policies that will be adhered to when protecting genetic data, usually through implementing security controls. While there are numerous controls in each of the National Institute of Standards and Technology (NIST) and International Standards Organization (ISO) documents, implementing all of the controls is not necessarily required, and the organization may need to make these standards into required policies, depending on their security posture.

When determining the use of such standards, one must consider the value or sensitivity of various forms of genetic data. The more valuable the data, the more security requirements are needed to protect the data. Selecting controls from either **NIST SP800-53** or **ISO/IEC 27002:2022-02** is recommended. In addition, other considerations for security requirements include information processing, storage, communications, legal, regulatory and contractual requirements.

Genetic data needs to be identified and inventoried by each institution and user in order to categorize the data based upon its defined attributes and characteristics. Following this categorization, the security requirements for the data are then known and the necessary controls can be implemented. From there, to maintain a robust security posture, regular planning, reviewing data categorization, and reassessing security gaps will be required.

BIO-ISAC shares recommendations for cyberbiosecurity requirements at **isac.bio**.

Genetic Data Categorization and Security Requirement Examples

The table below is a top-level categorization of major categories. Please refer to the implementation section below for key factors that will affect how specific uses may heighten or reduce the associated risk.

Examples from this table are not exhaustive, nor should they be taken to be immutable. As biological technologies evolve, so will the risk profiles. Reassessing the finer examples of categorization should occur on a yearly basis, at a minimum.

Category	Description	Data Examples	Minimum Security Requirements
High Risk	High risk of misuse/ high-value data Data from the high risk category will require advanced security measures to counter advanced persistent threats that could result in loss of valuable genetic data (i.e. intellectual property), or threats to biosecurity.	Human Whole Genome Data DNA of Dangerous Pathogens Whole genome prenatal sequencing	Require advanced strategies including restricted data connectivity or keeping the data offline and restricted user access through the highest authentication protocols. Twice yearly review of categorization and protections (minimum).
Medium Risk	Medium risk of misuse/ medium-value data Making specific security recommendations for this category can be challenging because of ambiguous security requirements. Labs that handle data from the medium risk category also must consider that the data that they generate and handle may change into the high risk category, so plans to increase security posture may be required.	DNA of common pathogens and/or their vectors Clinical Genetics (Not full genome data) Genome R&D (e.g. intellectual property, GMO research)	Require users to have MFA to access datasets and keycard access to facilities. Annual review of categorization and protections.
Low Risk	Low risk of misuse/ low-value data The low risk category will require basic security designed to protect against common cybersecurity threats. Data may eventually be made public, but a focus on operations and data integrity is still essential.	Genome sequences from readily available organisms Common research genomics (e.g. yeast, mouse, fruit fly)	Require basic login credentials and verification of users. Enable multi-factor authentication if possible. Annual review of categorization and protections.

Justification of Selected Categorization

Anonymized Human Genomics High Risk

Human genome sequences contain highly sensitive and personal information, including their susceptibility to certain diseases or behaviors. This data could be used to target specific populations or to gain unauthorized access to personal health data. Additionally, the genetic information obtained from full genome sequencing can be identified, as it may reveal familial relationships and ancestry.

DNA of Infectious Agents High Risk

The DNA sequence of dangerous pathogens, such as those on the list of **select agents** can be used to develop or refine bioweapons. Full genome sequences of these organisms have been shown in laboratory settings to fully recreate the organisms from synthetic DNA. While these capabilities are nascent, the risk profile of such capabilities is likely to increase rapidly as more sophisticated DNA synthesis technologies are developed and become more available.

Prenatal DNA Sequencing High Risk

This sequence data contains genetic information from an unborn child, a mother, as well as genetics from the pathogens that she may carry (i.e. her sexually transmittable diseases). Also, the combination of information from mother and child makes it easier to re-identify these data. If the child's genetic information were to be leaked, the child could be affected throughout its lifetime. Although this longer timeline results in a less precise threat landscape, we are confident the threat profile will continue to increase as time progresses and new technology develops.

DNA of Common Pathogens Medium Risk

Epidemiologists and virologists rely on DNA sequence data from common pathogens and their hosts/vectors to predict how diseases will spread and to develop countermeasures like vaccines. Ensuring availability and integrity of this data is required to monitor and control these diseases. These data may become more sensitive, depending on the virulence of the pathogens, the hosts they infect, or the rate of spread caused by the vectors.

Clinical Genetics Medium Risk

This data represents a security risk due to the sensitive nature of the data. Unauthorized access or disclosure can lead to privacy breaches and potential identity theft. Mishandling or exposure of Protected Health Information may also result in regulatory

non-compliance, legal consequences, and compromised patient trust. Importantly, if full genome data has been generated to produce the clinical genetic data, that full genome data may be PII and would belong in the high risk category. All forms of data generated and handled must be considered, not only the fully processed data. Similarly, careful consideration for data from other 'omics, such as from RNA sequencing is required to determine security categorization.

Genome Research & Development Medium Risk

The bioeconomy relies on genomic data to design and create new organisms. The full genome data of these organisms (e.g. intellectual property, GMO research) is sufficient to recreate them from synthetic DNA. This valuable intellectual property requires confidentiality and assurance of data integrity. Depending on the specific types of data inclusion by the user, this may require an escalation to High Risk categorization.

Common Research Genomics Low Risk

In general, this low risk genetic data may not require confidentiality. This data does, however, require integrity and availability due to its value to the research community. Because online systems are attractive targets for offenders, particularly when these systems are connected to operational assets (instruments, laboratories, data systems, etc.), baseline cybersecurity guidelines are strongly recommended for any mission critical and/or connected systems.

Implementation and Use

It is important to acknowledge that the implementation of this document will require periodic re-assessment, and tracking of modern vulnerabilities and risks. This field is in constant flux, with new vulnerabilities and threats developing weekly. While this may seem challenging, this guide will help you determine the interval at which you should consider re-evaluating risks and controls. Security norms and government regulations are also going to evolve over time. This guide is written with the end-use in mind, and focuses on the risk - thus advice derived from this document will last longer than specific controls tied prescriptively to specific file formats, software or DNA sequences. Any data that is held for long periods of time becomes a potential liability and awareness of the data held in databases is essential for determining the appropriate, current level of data security that is required. Continuous action planning can help determine the time and cost for additional security measures, helping to anticipate and facilitate security advances when they are required - ultimately avoiding costly breaches and sometimes irreparable damage to privacy.

SGenetic data must be categorized before or at the time of generation, and this categorization may change after it has been processed. Genetic data may be both sensitive and identifiable when it is generated (e.g. as whole genome data), but become less sensitive or more difficult to identify after it has been processed (e.g. into a diagnosis). Also, the quantity of data produced can affect its categorization. When planning, consider that certain genomic data may require a higher category simply because it is stored (E.g.: written to disk or stored in a database) or because of its accumulation into higher volumes or proximally to additional context data. Genetic databases that contain a high number of samples are more valuable to attackers than an isolated research record. However, just because it is a singular research record, it does not mean it is not valuable. If that research record is part of a wider-ranging campaign, you should consider the risk that results from the accumulation of these records. Leveraging this risk categorization and action planning approach should take into account the following criterion:

1. Misuse - Determine likely potential misuse during the pertinent data lifetime
2. Data Volume - Genetic data can require more security when aggregated.
3. Pathogenicity - High pathogenicity requires higher security and can change.
4. Reidentification of human data - Human genome data is identifiable, and other ‘omics data can also be identifiable, depending on its format.

A database may require additional security because the amount of data of one type has grown. A database with only a few full human genomes may be only a marginal target for any attacker, but if that database grows to more than a thousand individual human genomes, it becomes a greater risk and thus requires higher levels of security. Recent regulatory efforts call for new export controls that restrict sharing human genomic data with countries of concern only for data of a specific size with limits on genomic data being set for a thousand or more records.

Similarly, as the vulnerabilities and risks surrounding any type of data can evolve, so too can the combinations of different data types, or accumulation of any one type of data.

For example, research on genomics may only require a green (low) or amber (medium) categorization. If that genomic data is combined with pathogens or vectors that may signify genetic vulnerabilities to those pathogens or vectors, the red (highest) level of data security should be required. For example, agricultural genomic data may be of only low or medium security, but when combined with data on the yield of crops or livestock under stressed conditions and/or infection, the data could reveal sensitive vulnerabilities.

Conclusion

To truly support a safe, secure bioeconomy, industry must open a dialogue with itself about the data mobilizing much of its work.

Ultimately, while some genetic data obviously fits into one of these three categories, much of the data categorization for genetic data will require input from security experts as well as industry stakeholders. We recommend you check back annually and use this document to re-evaluate the categorization and risk matrix for the data you and/or your organization are responsible for. Remember to evaluate the risk based on the updated data volume, storage location, encryption parameters, and contextual data stored.

This document is meant to serve as an easy guide to define these categories and then provide security controls and recommendations that can be applied to labs that generate and handle genetic data. Different countries and different industries can always refine the specific types of data that they determine to exist in each category. From there, stakeholders can work to define the specific controls required for each category. As the risk classes and security controls become well defined, a generalized framework for genetic data categorization can be developed.

BIO-ISAC maintains free resources for securing facilities and assets and is available to support a confidential evaluation: https://isac.bio/device.

BIO-ISAC Genomic Data Categorization Workgroup

Authors

Sterling Sawaya, GeneInfoSec Inc.
Aaron Hanson, GeneInfoSec Inc.
Garrett Schumacher, GeneInfoSec Inc.
Terrion Fields, HudsonAlpha Institute for Biotechnology
Phillip Whitlow, HudsonAlpha Institute for Biotechnology
Scott Ross, HudsonAlpha Institute for Biotechnology

Contributors

Charles Fracchia, Black Mesa and BIO-ISAC cofounder and chairman of the board
David Molik, AgBioData Research Coordination Network
Kristina Zudock, Johns Hopkins University Applied Physics Laboratory
Whitney Zatzkin, Bioeconomy ISAC

bioisac
Creating Collaborative Threat Intelligence for the Bioeconomy

Additional Reading

Cybersecurity

Martin, N., Pulivarti, R., Wagner, J., Maragh, S., McDaniel, J., Zook, J., ... & Wojtyniak, M. (2023). *Cybersecurity Framework Profile for Genomic Data* (No. NIST Internal or Interagency Report (NISTIR) 8467 (Draft)). National Institute of Standards and Technology. https://csrc.nist.gov/News/2023/cybersecurity-of-genomic-data-nist-ir-8432

Pascoe, C. , Quinn, S. and Scarfone, K. (2024), The NIST Cybersecurity Framework (CSF) 2.0, NIST Cybersecurity White Papers (CSWP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.CSWP.29, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=957258

Ross, R. (2012), Guide for Conducting Risk Assessments, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.800-30r1

Schumacher, G. J., Sawaya, S., Nelson, D., & Hansen, A. J. (2020). Genetic information insecurity as state of the art. *Frontiers in bioengineering and biotechnology*, *8*, 591980. https://www.frontiersin.org/articles/10.3389/fbioe.2020.591980/full

Identifiability of DNA

Sero, D., Zaidi, A., Li, J., White, J. D., Zarzar, T. B. G., Marazita, M. L., ... & Claes, P. (2019). Facial recognition from DNA using face-to-DNA classifiers. *Nature communications*, *10*(1), 2557. https://www.nature.com/articles/s41467-019-10617-y

McGuire, A. L. (2008). Identifiability of DNA data: the need for consistent federal policy. *The American Journal of Bioethics*, *8*(10), 75-76. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771195/

Incidents

Spyscape, *Episode 48: The Spy in the Cornfield*, https://spyscape.com/podcast/the-spy-in-the-cornfield