A Review of Threat Vectors to DNA Sequencing Pipelines

Tyler Rector
Old Dominion University

A Review of Threat Vectors to DNA Sequencing Pipelines

Tyler Rector

Advised by Shobha Vatsa, Lecturer at Old Dominion University

COVA CCI Undergraduate Research, Fall 2023

Abstract

Bioinformatics is a steadily growing field that focuses on the intersection of biology with computer science. Tools and techniques developed within this field are quickly becoming fixtures in genomics, forensics, epidemiology, and bioengineering. The development and analysis of DNA sequencing and synthesis have enabled this significant rise in demand for bioinformatic tools. Notwithstanding, these bioinformatic tools have developed in a research context free of significant cybersecurity threats. With the significant growth of the field and the commercialization of genetic information, this is no longer the case. This paper examines the bioinformatic landscape through reviewing the biological and cybersecurity threats within the bioinformatics pipeline. It is found that there are significant security deficits within existing bioinformatic databases. Additionally, it is found that there is a theoretical trojan threat posed by unverified malicious DNA sequences.

1.1 Introduction to DNA Sequencing and Analysis

DNA is the primary molecule carrying genetic information. DNA is encoded using four nucleotides A, C, G, and T, and is analogous to any form of digital information. DNA genetic information typically follows a standardized pipeline that involves the preparation of DNA samples, sequencing of the DNA molecule, bioinformatic analysis, and storage of the genetic information. The Decrease in the cost of DNA sequencing has significantly outpaced Moore's Law, described as a trend that computing power doubles every two years, with the cost per genome estimated at $95,263,071 in September 2001 and $525 as of May 2022 [13]. The DNA sequencing landscape changes fast, especially with the development of technologies such as Oxford Nanopore sequencing [20]. The structure of how raw sequencing data is generated and analyzed will change as time goes on. However, some fundamentals of the DNA sequencing and analysis pipeline will continue to remain. These fundamentals in the DNA sequencing and analysis pipeline are vulnerable to cyber-attacks and pose a threat to the growing biotechnology industry. This paper hopes to review where those vulnerabilities in the DNA sequencing and analysis process are and propose solutions to mitigate them.

1.2 Challenges with Genomic Databases

Genomic data is a unique identifier of one's family history and disease risk along with being important tools in the development of biotechnologies and drugs. DNA can store massive amounts of data which is why it is being explored as a future data storage method [5]. But that conversely means that when analyzed the sheer amount of data necessitates the use of storage servers. This has necessitated the development of publicly available online databases such as GenBank to store the volumes of genomic data being generated [6]. But these genomic databases are at risk of the same adversarial attacks that any public server is susceptible to. Furthermore, genomic data has a unique set of challenges that complicate the attempts to secure genomic databases. Genomic databases have the same legal restrictions that any individualized medical information has in the United States. The federal law Health Insurance Portability and Accountability Act (HIPAA), is the main framework that protects Personal Health Information (PHI) but genetic information is also governed by the Genetic Information Nondiscrimination Act (GINA) which protects against discrimination, in an employment and insurance context, based upon genetic information [17, 9]. While genomic samples are constantly being shed by individuals every day, the risk of an adversarial collection is slim and the risk is localized to one individual. Genomic databases store the genetic information of multiple individuals who are likely to have an initiating factor for genomic analysis such as a disease, family history, work requirements, or criminal investigation. This is exacerbated by the problem that the exposure of one individual's genetic information can expose the likely genetic information of close relatives. So the effects of an adversarial attack on a genomic database are exponentially greater than the individual risk of an exposure of any one genome. Deidentification, a standard tool used for sensitive medical information, is less effective in the context of genetic information as genomes themselves can lead to the identification of the individual within a sample. As demonstrated by Homer et. al. within a complex genomic DNA mixture the team was able to identify individuals with an individual contribution of 0.1% of the total genomic DNA within a sample [10]. Gymrek et. al. demonstrated the capability to extract and recover the surnames from deidentified genomic sequences from publicly available databases [8]. The implications of reidentification from genomic data can compromise the trustworthiness of both public and private databases.

1.3 Vulnerabilities in Shared Databases

Genomic databases are susceptible to the same cybersecurity threats that any public database faces. The development within a research context and the previously high cost of DNA sequencing have shielded genomic databases from cybersecurity threats. However, the advent of direct to consumer companies such as 23&Me and the significant decrease in the cost of genomic sequencing has made these threats more salient. In 2019 the MIT Technology Review found that 26 million customers had uploaded their genetic data to commercial genomic databases [15]. The forensic genomics company Verogen had a security breach of their GEDmatch database [14]. Although no genomic data was compromised the Genomics company MyHeritage reported a mass email phishing attempt the next day possibly connected to the GEDmatch breach [2]. In an analysis of the bioinformatics tools and database used in the typical bioinformatics pipeline Arshad et. al. exposed multiple vulnerabilities and attack vectors. Some common features of the code in these tools were a lack of input validation, the use of obsolete functions, inadequate authentication, and the use of HTTP over HTTPS. This opened these tools and databases to potential cyberattacks such as denial of service attacks, attributed disclosure attacks, SGX attacks, buffer overflow, and homomorphic encryption attacks. Attributed disclosure attacks expose some of the unique vulnerabilities that genomic data carries. Publicly exposed Loci of a genome can be compared against anonymized genomic data to trace back to the individual from which it came [3]. Tao et. al evaluated common bioinformatic web applications and found similar results for the cybersecurity of these tools. The team found that 34.4% of these web applications were vulnerable to public exploit scripts and half of the analyzed websites used server software with high-risk vulnerabilities [16]. Within the field of pathogen genome research there are many unique challenges that make it uniquely vulnerable. An example is that high-risk genome sequences, such as Smallpox, are publically available to any anonymous user on NCBI [18]. Vinatzer et. al. discussed some of these unique vulnerabilities after examining multiple public pathogen databases. The analysis found that these databases had common vulnerabilities such as not requiring a username and password. The analysis also found that to prevent the compromisation of the genomic database an upload process has developed with curators overseeing the uploaded data. But the rapid growth in the field has reduced the effectiveness of this manual process. The analysis also found that the database hardware was a prime target for the mining of cryptocurrency and was often susceptible to SQL injection attacks [19].

1.4 Physical and Biological Threats

While database security is a pre-existing issue, with significant research, in cybersecurity the physical security side is more unique to the domain. DNA sequencing and the broader pipeline as a whole that involves the management of products ranging from samples to DNA Sequencers themselves. If any step is compromised along the way to the digitization of genetic information poses a more fundamental threat to the bioinformatics pipeline as a whole. Within the United States, DNA synthesizers are mandated to verify that the device is not synthesizing select agents and toxins such as SARS coronavirus (SARS-CoV) [1]. This principle could be extended to sequencers to detect a known list of sequenceable toxins, such as SARS-CoV, and place a tracer tag within the sequenced data. Peter Martin Ney suggests the detection of shellcode embedded within malicious DNA sequences should be completed prior to synthesis

[12]. If malware or data that can initiate preinstalled malware is sequenced it can compromise the integrity of all prior and future data collected. M.S Islam et. al. demonstrated the feasibility of synthesizing and detecting trojan sequences that can carry these forms of malware [11]. Ho Bae et. al. created a framework, using machine learning, for the detection of forms of steganography, the process of concealing information within data or physical objects, within malicious sequences [4]. Proper sample security plays an important role in the security of the initial bioinformatics pipeline. Especially for fields in which the integrity of DNA sequences is paramount such as forensics. Notwithstanding, Peter Martin Ney suggests it may be insufficient due to the presence of sample bleeding within the DNA sequencing process. The paper encourages the implementation of verification of sample sources and the use of techniques that minimize sample bleeding [12].

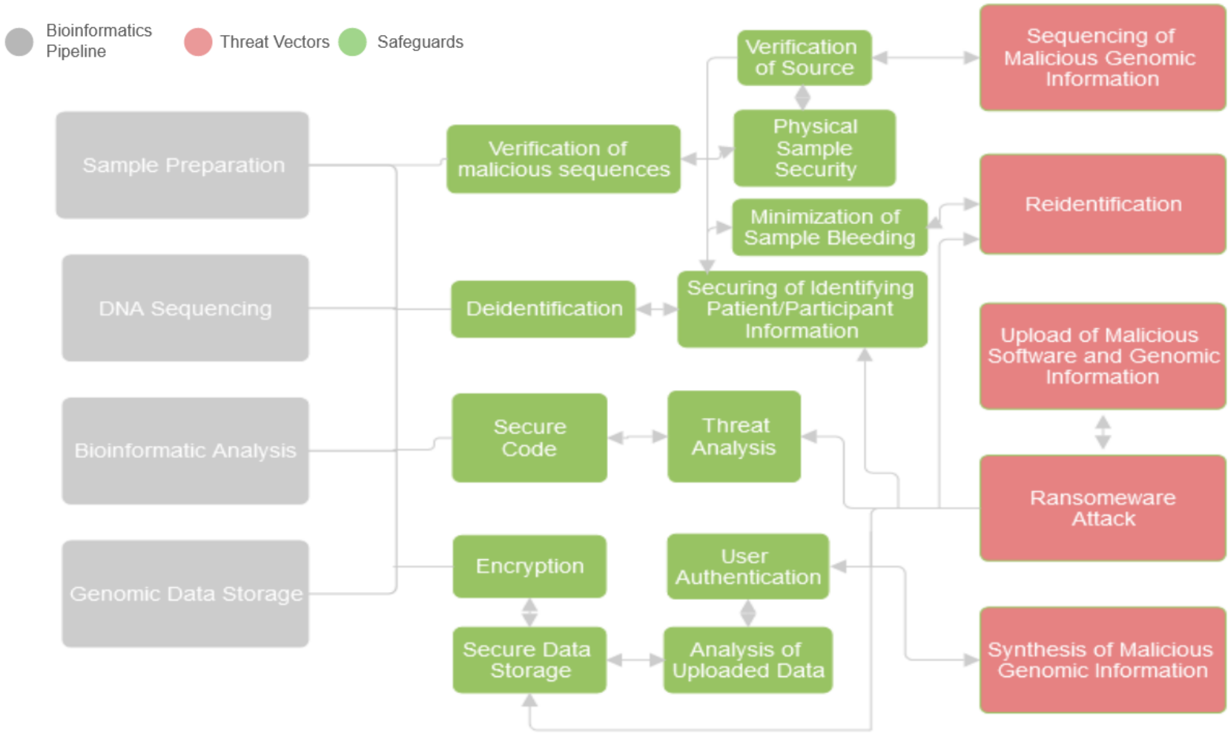

Figure 1: Threat Vectors to the Bioinformatics Pipeline and Safeguards

2.1 Future Directions

Most of the cybersecurity threats identified such as email phishing, SQL injection, and lack of verification are not unique problems. These issues necessitate the use of more diligence in the development of current and future bioinformatic software and toolsets. Additionally, more research is needed to investigate the vulnerabilities in bioinformatic software programmed in R. R ranks within the top 10 popularly used programming languages and is widely used in the field of bioinformatics [7]. Reidentification poses a unique threat to public databases and necessitates a more stringent verification of those who use these databases. Additionally, while malicious sequences pose a limited risk due to the strong technical knowledge required to develop such an attack, in the future, the proliferation of DNA technologies may make such attacks more feasible.

References

[1] *42 CFR 73.3 -- HHS select agents and toxins.* (n.d.). https://www.ecfr.gov/current/title-42/chapter-I/subchapter-F/part-73/section-73.3

[2] Admin. (2020, October 29). *Security alert: malicious phishing attempt detected, possibly connected to GEDmatch breach*. MyHeritage Blog. https://blog.myheritage.com/2020/07/security-alert-malicious-phishing-attempt-detected-possibly-connected-to-gedmatch-breach/

[3] Arshad, S., Arshad, J., Khan, M. M., & Parkinson, S. (2021). Analysis of security and privacy challenges for DNA-genomics applications and databases. *Journal of Biomedical Informatics*, *119*, 103815. https://doi.org/10.1016/j.jbi.2021.103815

[4] Bae, H., Min, S., Choi, H.-S., & Yoon, S. (n.d.). *DNA Privacy: Analyzing Malicious DNA Sequences Using Deep Neural Networks*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9170842

[5] Dong, Y., Sun, F., Ping, Z., Ouyang, Q., & Qian, L. (2020). DNA storage: research landscape and future prospects. *National Science Review*, *7*(6), 1092–1107. https://doi.org/10.1093/nsr/nwaa007

[6] GenBank. (2012). *Nucleic Acids Research*, *41*(D1). https://academic.oup.com/nar/article/41/D1/D36/1068219

[7] Giorgi, F. M., Ceraolo, C., & Mercatelli, D. (2022). The R language: an engine for bioinformatics and data science. *Life*, *12*(5), 648. https://doi.org/10.3390/life12050648

[8] Gymrek, M., McGuire, A. L., Golan, D. E., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321–324. https://doi.org/10.1126/science.1229566

[9] HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996. (1996). *GovInfo*. https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

[10] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using High-Density SNP genotyping microarrays. *PLOS Genetics*, *4*(8), e1000167. https://doi.org/10.1371/journal.pgen.1000167

[11] Islam, M. S., Ivanov, S., Awan, H., Drohan, J., Balasubramaniam, S., Coffey, L., Kidambi, S., & Srisa-An, W. (2022). Using deep learning to detect digitally encoded DNA trigger for Trojan malware in Bio-Cyber attacks. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-13700-5

[12] Ney, P. (2019). *Securing the Future of Biotechnology: A Study of Emerging Bio-Cyber Security Threats to DNA-Information Systems*. University of Washington.

[13] Nhgri. (2019, March 13). *DNA sequencing costs: data*. Genome.gov. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

[14] Osypian, T. (2020, July 21). *GEDMatch Incident Response*. Verogen - a QIAGEN Company. https://verogen.com/gedmatch-incident-response/

[15] Regalado, A. (2020, June 18). More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*.

https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/

[16] Tao, T., Chen, Y., Liu, B., Jin, X., Yan, M., & Ji, S. (2019). Security Analysis of Bioinformatics WEB application. In *Advances in intelligent systems and computing* (pp. 383–397). https://doi.org/10.1007/978-3-030-16946-6_30

[17] *The Genetic Information Nondiscrimination Act of 2008*. (n.d.). US EEOC. https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008

[18] *Variola virus, complete genome - Nucleotide - NCBI*. (n.d.). https://www.ncbi.nlm.nih.gov/nuccore/NC_001611.1

[19] Vinatzer, B. A., Heath, L. S., Almohri, H. M. J., Stulberg, M. J., Lowe, C. J., & Li, S. (2019). Cyberbiosecurity challenges of pathogen genome databases. *Frontiers in Bioengineering and Biotechnology*, *7*. https://doi.org/10.3389/fbioe.2019.00106

[20] Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x