Medical and Biological Data Security

Rebecca Newman

1 Abstract

A strong emphasis is placed on safety and health in medical and biological reasearch. Why, then, does data security often get overlooked? Though the monetary value of this data as a payload is not as high as other exploits, attackers with specific interests may target this data for a variety of reasons. Assuming immunity due to attacker disinterest is a dangerous and ignorant attitude. With hospital and lab equipment connected to the internet and sensitive information stored in shared databases for collaborative research, cyber security must be a priority for the researchers themselves.

Manipulation of the output from this equipment, such as Short Tandem Repeat(STR) counts from next-generation sequencing technologies, though difficult, could have malicious outcomes such as tampering with evidence in criminal investigations. More minor incident are also possible, such as buffer overflow attacks from using open source pipelines to analyze reads.

2 Introduction

Medical researchers are quite aware of the risks that their research could pose to public heath and safety. Those who work closely with this kind of data are subjected to trainings, lecutres, and certifications relating to protection of human life and anonymity of data. However, there is a distinct lack of focus on the vulnerability of this data with regard to cyber security. Many of the machines that produce and store results from these experiments are connected to unsecure networks and analysis software is not written with security in mind. An exerpt from a Q&A on the Illumina cloud storage blog shows a nonchalance towards the security of sensitive lab and medical data stored using Amazon Web Services:

> "**Isnt a big public cloud provider a huge target, and so inevitably vulnerable to attack?**

> Obviously a criminal attacker that finds a vulnerability isnt going to tell AWS about it. But in the words of the famous cartoon I dont have to outrun that bear, I only have to outrun you: if someone breaks into Amazon their target will almost certainly be easily monetized data such as credit card numbers, not genomic data." [1]

The naivete of this statement is rather jarring, especially when noting that the statement comes from one of the largest distributors of next-generation sequencing technologies. When one understands the value of medical data, especially if that data is being used for something such as DNA forensics, it is not out of the realm of possibility that obtaining it would be an attacker's goal.

3 To The Community

The motivation for this paper is to urge those who handle sensitive biological and medical data to educate themselves on the possible exploits of thier data as well as the risks to and vulnerabilities of their pipelines. Throughout this paper I use next-generation sequencing as my example technology, but this is by far and away not the only system vulnerable to attack. A stronger emphasis must be placed on security in the biological sciences, and the pervasive attitude of "it wouldn't happen to me" must be abandoned.

4 Next-Generation Sequencing Technology

From the introduction of Sanger sequencing in 1977, the use of DNA as a tool in biological, diagnostic, and forensic sciences has exploded. The cost to sequence DNA has also dropped significantly from that time. The current industry standard for sequencing has shifted to Next-Gen technologies, which are able to process millions of sequence reads at the same time[2]. These technologies are prevalent now in many hospitals and labs across the US. As these machines are often connected to the internet or internal hospital networks for storage of results in databases, they remain vunerable to attack. Shodan even reveals that there are exposed and vulnerable sequencing technologies and databases storing their results[3]. In addition, universities and labs that share this data for collaborative research often use FTP

[1] https://blog.basespace.illumina.com/2011/12/13/basespace-security/

[2] Elaine R. Mardis, The impact of next-generation sequencing technology on genetics, In Trends in Genetics, Volume 24, Issue 3, 2008, Pages 133-141, ISSN 0168-9525, https://doi.org/10.1016/j.tig.2007.12.007. (http://www.sciencedirect.com/science/article/pii/S0168952508000231)

[3] https://www.shodan.io/host/128.249.176.22, https://www.shodan.io/host/128.223.20.127

as the protocol for sharing these files. Many FTP portals associated with US research institutions are also exposed on Shodan.

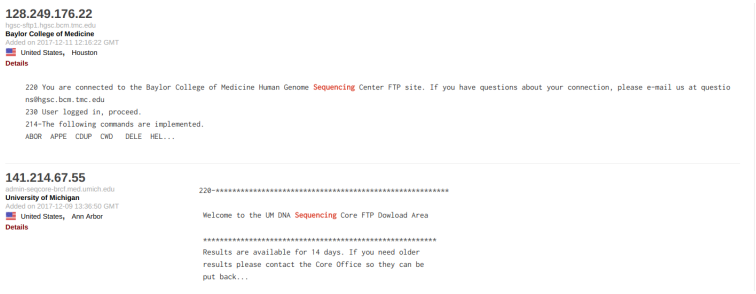


Figure 1: Exposed FTP sites associated with universities containing sequencing data

Taking advantage of the vulnerabilities in high throughput sequencing technologies and their data as they relate to DNA forensics can lead to detrimental consequences. With results from such equipment being used as evidence in criminal cases, provided that the lab followed regulations during DNA processing, modification of the output from these machines could convict or exonerate those charged with high profile crimes[4]. A less malicious exploit could be to change the results of experiments such as paternaty tests and clinical trials for monetary gain in civil proceedings.

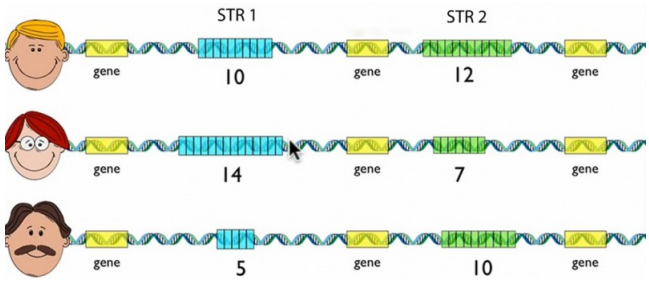
Consider the following example. Currently, it is standard to use Short Tandem Repeats (STRs) as a tool for human identification. STRs are long strands of repeated nucleotide units at one of 13 specific loci in the human genome. While the loci remain constant throughout the population, the number of repeats varies by individual[5]. Due to the fact that having a certain number of repeats on one loci is independent from another loci, the theoretical probability that two people share the same set of STRs is remarkably low. Therefore, this approach is considered highly accurate in indentifying matches between two DNA samples[6]. Though STR analysis has historically been conducted by Polymerase Chain Reactions (PCR) and Capillary Electrophoresis (CE), it has been moving to Next-Gen Sequencing as it becomes

[4]National Research Council (US) Committee on DNA Technology in Forensic Science. DNA Technology in Forensic Science. Washington (DC): National Academies Press (US); 1992. 6, Use of DNA Information in the Legal System.

[5]Bornman DM, Hester ME, Schuetter JM, Kasoji MD, Minard-Smith A, Barden CA, et al. Short-read high-throughput sequencing technology for STR genotyping. BioTechniques (2012)

[6]Weir BS. The Rarity of DNA Profiles. The annals of applied statistics. 2007;1(2):358-370. doi:10.1214/07-AOAS128.

less cost-prohibitive.



[7]

Figure 2: A visual representation of the variation in STRs in the human population.

A sample workflow for the processing of this kind of data is as follows[8]: A library is created from the DNA samples by "marking" where on the DNA strand the reads are to begin and end. These are called fragments. In STR processing, the desired loci is marked. These strands are then copied and duplicated many times by PCR. Then, the copied fragments are clustered based on which original strand they came from. Next the clusters are sequenced using the sequencing technologies. Finally, the data is uploaded and processed via bioinformatics pipelines. It is in these last two steps that a malicious actor can obtain or modify data in a stealthy way. By modifying the number of repeats a machine reads, or the read statistics, an individual can be misindentified. A man in the middle attack can replace the contents of the output file or spoof packets being sent to a database to contain desired information. The other point of possible exploit would be during the analysis phase. Data is often shared among researchers and analyzed with open source software tools. These tools are not always written with security in mind.

5 Man-in-the-Middle Attacks

A man in the middle attack occurs when a malicious party embeds themselves in a system in such a way that the client believes the malicious party is the server, and the server believes the malicious party is the client. This allows the attacker to gain access to the information that the client was exchanging with the server [9]. Once this arrangement has been set up, the attacker can inject their own information into the packets being sent between them. These attacks are hard to detect, and are often hard to prevent from the client side.

[7] http://www.healforce.com/en/index.php?ac=article&at=read&did=439
[8] https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
[9] https://www.veracode.com/security/man-middle-attack

Figure 3: The left image depicts normal communications between a miSeq sequencer and a router, while the right shows the flow of communication during a man-in-the-middle attack.

One of the common methods of carrying out this attack is ARP poisoning. This is accomplished by sending fraudulent ARP data over a network, which leads to the linking of the IP address of a server with the malicious person's MAC address[10]. One of two things could happen next. Either the traffic from the machine could be altogether stopped by the attacker, or the results could be modified. If the attacker has access to falsified read files, they could simply send those files as the traffic, instead of changing packet information. When the packet modification approach taken, the files in transit from the lab device to the storage device can be intercepted and changed. *Netsed* is one such open source tool that allows for this to happen to TCP and UDP packets[11]. This tool sets up a transparent proxy to redirect traffic from the client to a specific place, wherein the contents of the incoming packets can be modified. The traffic is then forwarded to the server. Specified replacement commands allow all instances of a string found in the packets to be replaced by another value. There are also options to replace anything found between two strings with another string. For sequencing this is particularly useful, as sequencer output files often have the format: "$>$*seq_id* sequence" and are stored in plain text files.

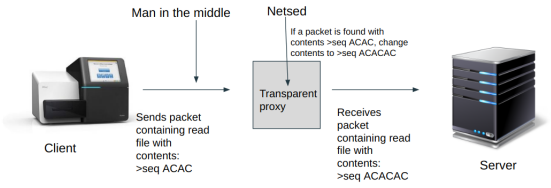


Figure 4: Theoretical setup of packet modification attack.

[10]https://www.veracode.com/security/arp-spoofing
[11]https://github.com/xlab/netsed/blob/master/README

6 Buffer Overflow

The OWASP website describes a buffer overflow as follows:

> "A buffer overflow condition exists when a program attempts to put more data in a buffer than it can hold or when a program attempts to put data in a memory area past a buffer...Writing outside the bounds of a block of allocated memory can corrupt data, crash the program, or cause the execution of malicious code[13]."

In C and C++, the scan, strcpy, scanf, and get function calls are all particularly vulnerable to buffer overflow attacks. A 2017 paper found that there were instances of these function calls in many of the common open-source analysis pipeline software[14]. Some of this software is even included with Illumina machinery. They found that Next-Gen Sequencing pipeline software contained on average 2.005 calls to the aforementioned functions per 1000 lines of code. Next, this was compare to a set of control software found to have 0.185 calls per 1000 lines of code. This was found to be a significant difference (p=0.027). The authors of the paper even contacted the software developers to share their findings, who responded that they had not thought extensively about the security impact of their code.

Using the previous example of modification to sequencing data, it is possible to inject code that changes the way sequences are analyzed. Even minor adjustments to the statistical calculations can have a major impact on lab results. Should an adversary wish to change experiment results, exploiting such vulnerabilites in common pipelines is a simple and effective method.

7 Action Items

There are many ways to mitigate the potential risks surrounding biological and medical data. The FTP protocol should not be used to send and recieve sensitive biological information. Also, it can be helpful to encrypt data; sequencing data should not be sent in plain text. When developing new software tools, it is important to check them through a program like Veracode, which finds potential vulnerabilities. It is also important to be familiar with known vulnerabilities for the chosen programming language.

[13] https://www.owasp.org/index.php/Buffer_Overflow

[14] P. Ney, et al. Computer Security, Privacy, and DNA Sequencing: Compromising Computers with Synthesized DNA, Privacy Leaks, and More (2017) https://dnasec.cs.washington.edu/dnasec.pdf

8 Conclusion

While the targeting of medical and laboratory data is very uncommon, it is imperative that researchers be prepared for adversarial pressure. This includes designing biomedical software with security in mind, understanding the flow of data through their respective pipelines, and maintaining software and network updates. The example of sequencing is just that - an example. There are many reasons why an adversary would target this kind of data, many of which are profitable. Cyber security in the biomedical sciences should be approached with the same gravity as physical health and safety because it has the potential to affect those outcomes as severely.