Digital DNA Lifecycle Security and Privacy: An Overview

Alsaffar[1], M., Hasan[1], M., McStay[2], G.P., Sedky[1], M.

[1] Department of Computing, AI and Robotics, School of Digital, Technologies and Arts, Staffordshire University, College Road, ST4 2DE, Staffordshire, United Kingdom.

[2] Department of Biological Sciences, School of Health, Science and Wellbeing, Staffordshire University, College Road, Stoke-on-Trent, Staffordshire, ST4 2DE, United Kingdom.

Corresponding author muhalb.alsaffar@research.staffs.ac.uk

Key points

- The digital DNA life cycle describes all the processes and usages once the DNA has been sequenced.
- One's privacy is threatened if their anonymised DNA is leaked; the threat level can be as high as creating somebody's face image.
- Attempts to secure genomic data can fail or may not always scale to cover actual life data.
- A new approach powered by a Machine Learning (ML) solution is required to protect genomic data.

Keywords:

Digital DNA lifecycle, Genomic privacy, DNA privacy, Genomic security, DNA security, Digital DNA attacks, DNA attack, Genomic privacy-preserving techniques, Direct-To-Consumer (DTC) security, recreational Genomics security, Genomics attacks.

Abstract

DNA sequencing technologies have advanced significantly in the last few years leading to advancements in biomedical research which has improved personalised medicine and the discovery of new treatments for diseases. Sequencing technology advancement has also reduced the cost of DNA sequencing, which has led to the rise of Direct-To-Consumer (DTC) sequencing e.g. 23andme.com, ancestry.co.uk etc. In the meantime, concerns have emerged over privacy and security in collecting, handling, analysing, and sharing DNA and genomic data.

DNA data is unique and can be used to identify individuals. Moreover, this data provides information on people's current disease status and disposition e.g. mental health or susceptibility for developing cancer. DNA privacy violation does not only affect the owner but also affects their close consanguinity due to its hereditary nature.

This paper introduces and defines the term 'Digital DNA Lifecycle' and presents an overview of privacy and security threats and their mitigation techniques for pre-digital DNA and throughout the digital DNA life cycle. It covers DNA sequencing hardware, software and DNA sequence pipeline in addition to common privacy attacks and their countermeasures when DNA digital data is stored, queried, or shared. Likewise, the paper examines DTC genomic sequencing privacy and security.

1. Introduction

DNA and genomic data security is vital to one's privacy. It can uniquely identify the owner and contains information about the individual's disposition to numerous diseases such as Alzheimer's and the likelihood of developing others e.g. mental disorders or other phenotypic traits [1]. Moreover, genomic data disclosure is not limited to a fixed period and does not only involve the owner. Due to the hereditary nature of the DNA, an adversary obtaining a target's genomic data can also predict a wide range of relevant traits to their close relatives and future descendants [2].

Genomic security is vital; if an adversary manages to gather one's genomic information, the adversary would then be able to predict phenotypes such as facial structures. The ability to predict physical traits and demographic information based on whole-genome sequences using Machine Learning (ML) has advanced over the years [3]. Physical traits prediction is a significant threat to privacy, and it also has important legal and ethical implications. The ability to predict physical traits will also affect the suitability of current informed consent, the practicality and value of de-identification of the supporting genomic information e.g. genomic owner's name and address [4]. Predicting facial structures based on whole-genome sequences has advanced even further. Research by Qiao et al. [5] demonstrated that facial characteristics such as cheeks, mouth shape and other facial features are related to as few as six genes and can be predicted from genomic data. Richmond et al. [6] give a brief overview of the various facial genetics variants that influence facial phenotypes.

There are many threats to one's privacy if the genomic information falls into the wrong hands. Genetic blackmailing is one of the main concerns. An adversary could identify individuals by

combining websites such as peoplefinders.com and publicly available (even though anonymised) genomic data from sources such as 23andme.com [7].

Genomic Discrimination (GD) is another concern as highlighted by Joly et al. [8]. The authors emphasised that there is no standard global approach to tackle GD. Many countries do not protect against GD, and approaches in countries that passed legislation to protect against GD suffer from many limitations such as the lack of public visibility, restrictive and non-flexible approaches with narrow protection (for example, the protection does not cover life insurance or travel insurance) and these legislations contain complex procedures.

These risks also affect the DNA data owner's kin due to correlation. Humbert et al. [9] demonstrated a novel reconstruction attack to infer the genomic data of individuals based on the genotype of their relatives which was achieved by using statistics in combination with Mendel's hereditary laws.

Despite privacy risks, genomic research is vital to improving human health such as applying translational genomic discoveries into clinical settings that enables the development of tailored interventions and the design of prophylactic approaches [10]. The use of the DNA and genomic data are also crucial for forensics and criminal investigations [11], paternity [12] and prenatal testing [13].

In recent years many reviews have been published regarding genomic security and privacy. These reviews generally tackle a specific issue or part of the overall digital DNA sequencing and usage such as privacy and privacy-preserving solutions for DNA sequence alignment and querying [14], [15], storing, sharing genomic data privacy and privacy-enhancing technologies [16], [17], [18], regulatory framework and consent [19], privacy while using the Cloud Computing [20], classification of genomic data privacy attacks and privacy-preserving solutions [21], [22],privacy-preserving techniques for genomic data [23] and review to Physical DNA sample security and digital DNA privacy [24]. To the authors' knowledge, no prior work has been presented as an overview for genomic security and privacy that covers the digital DNA security and privacy for pre-and post-DNA sequencing and DTC genomic testing.

This article contributes an overview of privacy and security of the physical DNA, hardware and software security used for DNA sequencing and genomic sequencing and usage processes. It discusses some of the latest literature on how the current methods employed to anonymise the DNA are insufficient to prevent individuals from being identified. It explores the privacy vulnerabilities and their current countermeasures in sequencing hardware and software. The paper introduces the term digital DNA lifecycle to encapsulate all the steps that follow the output of the DNA sequencers such as sequencing pipeline, genomic data querying, and sharing. It also reviews the vulnerabilities within DTC DNA testing and finally draws conclusions based on the information presented.

This paper is structured as follows. Section 2 introduces the concept of digital DNA lifecycle where the authors identify the legitimate access and the steps/phases for possible threats. In section 3, security vulnerabilities for the DNA Sequencing process and their countermeasures are discussed. Section 4 focuses on post-sequencing privacy vulnerabilities and their countermeasures. Section 5 highlights vulnerabilities associated with querying and sharing DNA and genomic data as well as common DTC vulnerabilities and methods used to protect the genomic data are examined. Finally, section 6 draws some conclusions based on the information presented in the previous sections.

2. Digital DNA lifecycle

DNA is a double helix structure that contains genetic information encoded as a sequence of building blocks called nucleotides [9]. The whole human genome consists of 3.2 billion base pairs. Over 99.9% of the genome is identical between two individuals. The remaining 0.1% is the variation that can be in the form of single nucleotide changes i.e. Single Nucleotide Polymorphisms (SNPs) along with insertions, deletions, inversions and translocations. This variation leads to the presence of alleles, variants of a locus (a sequence at an exact unique location in the genome) that are responsible for particular traits and phenotypes. However, as the human genome is diploid, most low loci are biallelic where Loki can take two possible alleles [25].

DNA sequence building blocks correlate to each other e.g. the presence of a specific nucleotide sequence in a particular location indicates and correlates to another nucleotide sequence presence in another location. This correlation is called Linkage Disequilibrium (LD) [26] which will be considered in one of the methods to secure genomic data in section 4.

The digital DNA lifecycle starts with DNA sequencing which requires a patient or customer to provide a sample (saliva, blood or hair etc.) to a clinic or a DTC organisation. DNA is extracted and sent to a DNA sequencing lab as shown in Figure 1. DNA is prepared for sequencing; then sequencers are used to sequence DNA where the output is generally presented in Sequence Alignment Map (SAM) format which is transformed to more usable forms via software. The output from this software is digital DNA files i.e. assembled digital DNA (using resources such as Ensembl [27]). The next step in the digital DNA lifecycle is to align the software output files to a reference genome. Once the DNA has been aligned, it can be saved on a storage account (local or remote) where a primary analysis could be performed or a variant file could be extracted.

The digital DNA file can also be shared with other organisations where secondary analyses could be carried out such as functional genomics which helps researchers answer some questions, for example, quantifying the correlation between polymorphisms and complex diseases such as cancer. This type of research relies on secondary or tertiary analysis and data sharing [28].

During the digital DNA life cycle, DNA and digital DNA are accessed legitimately by multiple groups of people such as lab technician, scientific researcher, IT personnel who maintains the infrastructure or the software used for DNA analysis etc. who need and have the right to access and work on the DNA sample. However, Digital DNA privacy is exposed to every stage such as the risk of trojans and malware infecting DNA sequencers or infrastructure to leak information. There are also flaws within DNA sequencing software that can be exploited to allow arbitrary code executions. DNA sequences privacy can be unmasked by an attacker while clinicians or researchers are querying or sharing digital DNA data. An attacker achieves this by using data aggregation, correlation, likelihood ratio or linkage attacks etc. There are also threats originating from DTC genomic testing where the privacy of

the DNA is at risk from carefully constructed queries submitted to these sites and vulnerabilities of DTC websites themselves.

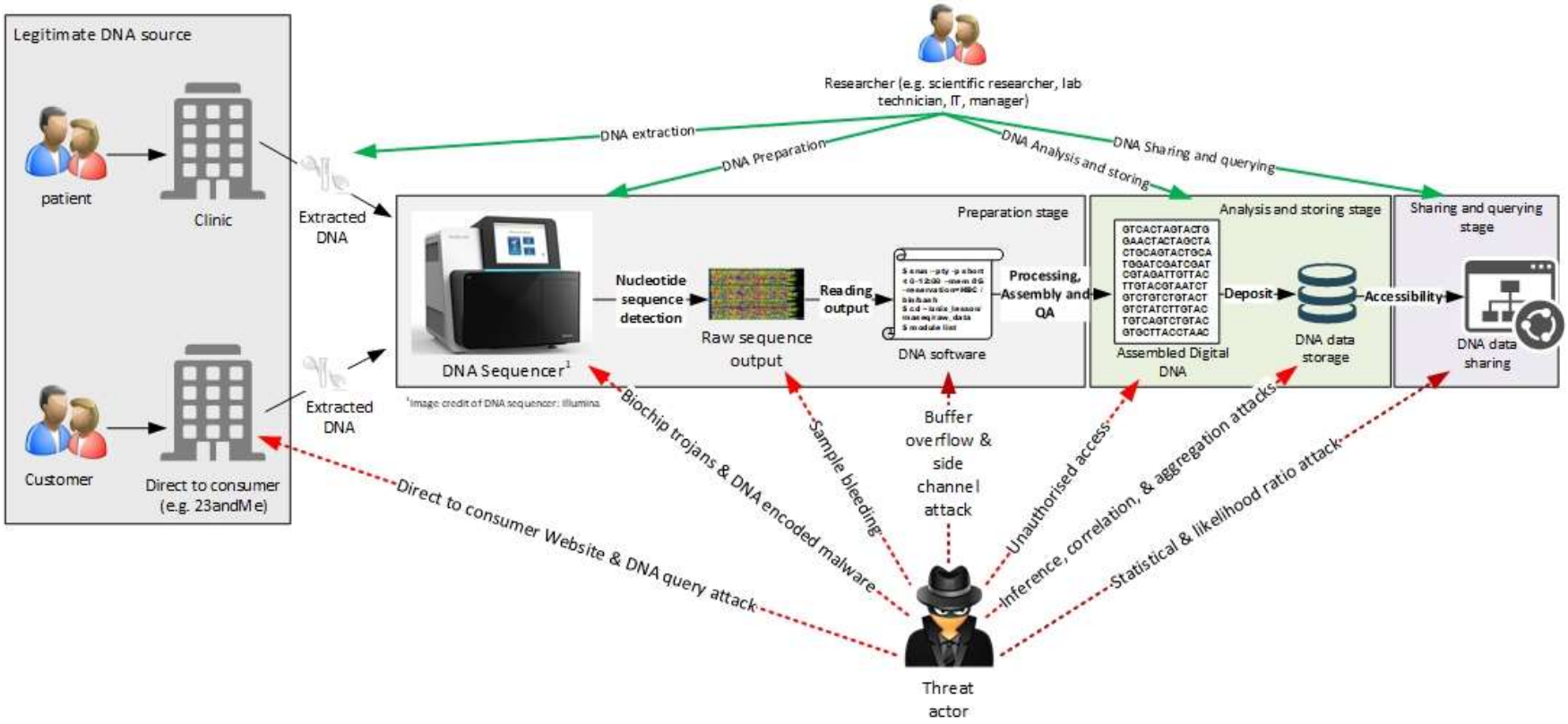

Figure 1. Digital DNA life cycle

DNA privacy vulnerabilities summary, where the patient has their DNA extracted, or a customer sends a saliva sample to a direct to consumer lab. The researcher refers to anybody with a legitimate need and has the right to access and work on the DNA sample (this can be a lab technician, scientific researcher, IT personnel who maintains the infrastructure or software used for the DNA analysis etc.). The figure shows that Human DNA is vulnerable at every stage where a threat actor can attempt to view or gain unauthorised access to that user's DNA.

3. Preparation stage vulnerabilities

3.1. Encoding malware in a strand of DNA

An active research area into molecular computing has shown that digital data can be encoded into a synthetic strand of DNA. Synthesised DNA is commercially manufactured using phosphoramidite chemistry [29].

3.1.1. Problem domain

Ney et al. [30] demonstrated an adversary's ability to encode a malicious computer code into a synthesised DNA sample. The authors were able to exploit a feature within the Linux operating system which allowed them to receive a copy of all the network traffic generated in the DNA alignment computer as shown in Figure 2. Even though the experiment was unreliable since the sequence reads were not 100% accurate, this implied that DNA could encode a malicious code.

3.1.2. Available solutions

The risk of DNA based attacks can be mitigated by ensuring sample source i.e. close monitoring of the biological sample from collection through sequencing. Besides, there are already regulations to prevent the synthesis of known dangerous DNAs such as synthesising harmful viruses [30] which could also be applied to a malicious code. However, sometimes it is not possible to trace the synthesised sample's origins because some biotech companies want to keep some sequence information confidential to protect their intellectual properties. Gallegoset al. [31] developed a method to create a digital signature for molecules of DNA to confirm the sample integrity, identity and to establish authorship with robustness to handle minor mutations.

In 2009, several of the largest DNA synthesis companies joined together to form the International Gene Synthesis Consortium (IGSC). IGSC developed the Harmonized Screening Protocol which offers practical guidance on implementing a safe DNA synthesis protocol. IGSC also created a Regulated Pathogen Database (RPD) which contains sequences and organisms subject to regulatory control or licensing. It published instructions on screening any requested synthesis against their RPD [32].

Even though it is possible to include harmful, malicious code into the IGSC database or enforce it through regulations, this has not been done yet. Researchers have yet to further explore this subject to determine how viable it is to create such an exploit.

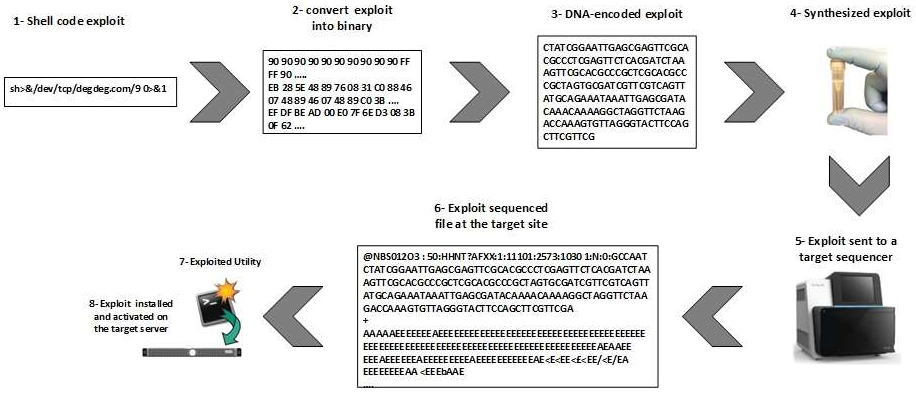

Figure 2. DNA encoded malware adapted from [30]

3.2. DNA sequencing hardware

The security of the DNA sequencers and downstream hardware is essential for data integrity.

3.2.1. Problem domain

Ali et al. [33] examined the vulnerabilities in the digital microfluidic biochip's supply chain. The microfluidic biochip can be used in a DNA sequencer hardware. The researchers identified that malware e.g. trojans could infect microfluidic biochips used in DNA sequencers, and the trojan is then used in various ways such as leaking or modifying information. Fayans et al. [34] pointed out that it is common for staff to use their office hardware for personal use, hence, increasing the chance of picking up malware that can only target and infect medical types of equipment that are not as protected as standard IT equipment. The researchers also point out the possibility sequencing machine can also be compromised at the time of manufacturing.

Another vulnerability to DNA privacy stems from DNA sequencers hardware sequencing technology, as DNA sequencers commonly sequence thousands of samples from different sources simultaneously; this technique is known as multiplexing. Multiplexing relies on assigning unique 6 to 8 digit identifiers to each sample; these identities can then be used to identify the sample during the demultiplexing process. The demultiplexing process (which is the process of separating the samples

from each other) is not perfect; a wrong DNA sequence could be assigned to the incorrect identifier which is known as sample bleeding [30]. Sample bleeding commonly exceeds 1% on some widely used sequencing platforms [35].

Ney et al. [30] demonstrated that multiplexing could be used as a side-channel attack to sabotage or influence a sequencing run or reveal information about the sample itself.

3.2.2. Available solutions

Ali et al. [33] suggested several methods to improve the security of microfluidic biochips such as using the digital watermark or utilising code analysis at the actuation sequences to detect if trojans are inserted. Ney et al. [30] suggested assigning two identifiers to the sample instead of one or altering the algorithm used e.g. Long Template Protocol [36] to minimise sample bleeding.

3.3. DNA sequencing software

DNA sequencing software is another significant part of the DNA sequencing pipeline as the DNA sequencers initial output is rarely usable; meaningful data is usually obtained from downstream processing and analysis. These downstream processes are typically carried out in stages where the end of each step feeds into the start of the next one [37].

3.3.1. Problem domain

Most of these programs are written by small research groups and might not have been subjected to software security scrutiny. Many software used in the downstream process are written in C, C++ and Java. These languages are known to be vulnerable to a buffer overflow flaw [37]. Fayans et al. [34] highlight that vulnerabilities with the genomic software could be exploited to gain unauthorised access to the computer or network resources and can also be used to leak information, crash or disrupt various services, especially if the software is running with higher privileges.

Ney et al. [30] assessed a sample software covering every stage of the DNA downstream pipeline. The sample was grouped into specific categories and found to use many known insecure functions/commands such as “strcpy” as shown in Table 1.

3.3.2. Available solutions

Ney et al. [30] suggested that software security can be improved by following software security best practices including regular patching and updates.

Table 1: Sample software which is used in DNA analysis that was found to have insecure function call or static buffer declaration, the number has been normalised by the number of appearance to 1000 lines of code [30]

Category	Program	Version	Lines of Code	Normalised Count (Total Count)					
				strcat	strcpy	sprintf	vsprintf	gets	static buffers
NGS Analysis									
Preprocessing	fastx-toolkit	0.0.14	3189	0.314 (1)	0.314 (1)	0 (0)	0 (0)	0 (0)	14.425 (46)
	fqzcomp	4.6	2066	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	23.223 (48)
Alignment	bowtie2	2.2.9	58377	0 (0)	0 (0)	0 (0)	0 (0)	0.017(1)	3.272 (191)
	bwa	0.7.15	13496	1.926 (26)	2.223 (30)	0.222 (3)	0 (0)	0 (0)	10.966 (148)
	hisat2	2.0.5	80930	0 (0)	0 (0)	0 (0)	0 (0)	0.012(1)	2.508 (203)
	STAR	2.5.2b	14760	0 (0)	0.136 (2)	0.271 (4)	0 (0)	0 (0)	3.388 (50)
De novo assembly	MIRA	4.0.2	69,853	0.014 (1)	0.115 (8)	0.115 (8)	0 (0)	0 (0)	1.904 (133)
	velvet	1.2.10	22,794	1.228 (28)	2.106 (48)	1.185 (27)	0 (0)	0 (0)	2.588 (59)
	SOAPdenovo2	2.04-r240	37,010	0 (0)	0.351 (13)	3.161 (117)	0 (0)	0 (0)	4.945 (183)
Alignment processing	samtools	1.5	56,979	0.351 (20)	0.228 (13)	0.509 (29)	0 (0)	0 (0)	30247 (185)
	bcftools	1.5	77,707	0.090 (7)	0.283 (22)	0.360 (28)	0 (0)	0(0)	4.375 (340)
RNA-seq	cufflinks	2.2.1	68,539	0.058 (4)	0.817 (56)	1.984 (136)	0.029 (2)	0 (0)	4.844 (332)
ChIP-seq	PeakSeq	1.3	6,806	0.147 (1)	3.967 (27)	3.526 (24)	0 (0)	0 (0)	7.787 (53)

3.4. Summary

DNA Sequencing hardware and software is vulnerable to misuse and errors (unintentional or otherwise). Figure 3 shows a summary of preparation stage vulnerabilities which consists of three levels; the top layer is the vulnerability source, the middle layer describes the attack vector and the bottom layer demonstrates the methods used to mitigate or reduce the risk of the attack vector.

Extracted DNA sources can be tampered with to disrupt the sequencing cycle or create malware e.g. a worm that can infect the downstream computers and allow the attacker to receive a copy of their network communication. To reduce the risk, it is important to ensure that the DNA source is trusted and tracked which can be achieved by digitally signing DNA molecules.

Two vulnerabilities reside in the DNA sequencer hardware i.e. trojans which can infect sequencer hardware, and the sequencers multiplexing flaw (sample bleed). Both flaws allow the attacker to infer or influence the sequenced samples. To reduce the risk of trojans, sequencers boot sequence

check to ensure the boot code have not been modified. And to use multiple identifiers to minimise the effect of the sample bleed.

DNA sequencing software is another vulnerability source in the preparation stage where insecure function calls within the software can cause side-channel attacks or allow the attacker arbitrary code execution. Software security best practice guidelines should be used to mitigate and reduce the risk.

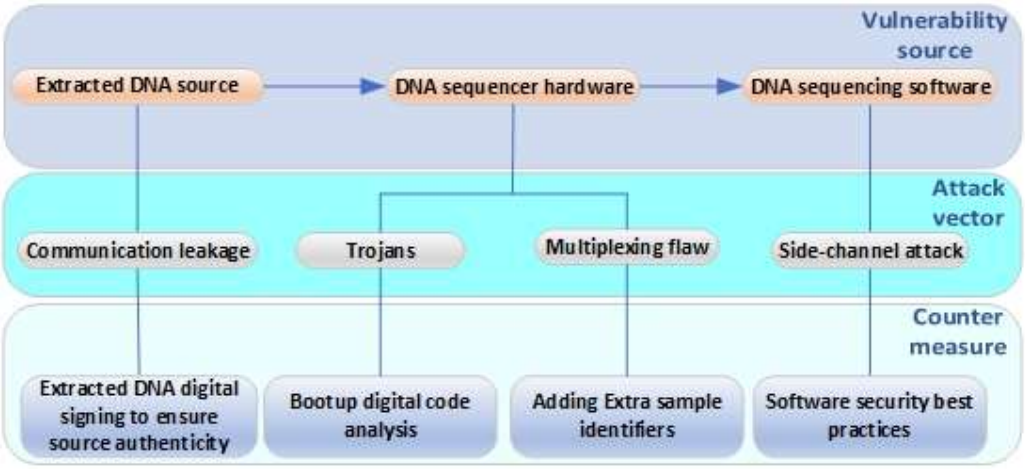

Figure 3. Summary of vulnerabilities associated with the preparation stage and their countermeasures

4. Analysis and storing stage vulnerabilities

4.1. DNA sequence read

DNA Sequence read lengths depend on the sequencer's model or technology and newer sequencer models tend to produce longer DNA read segments. Over the past few years, many methods for securing these reads have been developed. However, these methods have been mainly for short reads and have become less effective in protecting DNA with long read segments [38].

4.1.1. Current solutions

Cogo et al. [39] introduced a technique to classify and split DNA sequence reads to either privacy-sensitive or non-sensitive sections depending on which criteria they meet based on the reference knowledge database. Decouchant et al. [38] introduced a new method to secure DNA reads using

the bloom filter-based approach to identify sensitive reads. This approach tests if the reads are part of a previously built dictionary of known sensitive reads.

Fernandes et al. [40] introduced a novel method built on the existing bloom filter to classify the read data into sensitive and non-sensitive reads. The approach presents multiple levels of sensitivity classifications and access. Suppose an adversary managed to mount an attack and gain access to one partition of the sequence reads within a given sensitivity level. In that case, the adversary will not infer any more sensitive data from the other parts due to different access requirements. Gholami et al. [41] proposed separating the reading stage from the concatenation of the DNA fragments stage which happens within the DNA sequencer. The proposal is to outsource and distribute the reading stage and add ambiguity to prevent unauthorised assembly at the outsourced service.

4.1.2. Critical analysis

Hasan et al. [42] argued that using a pre-defined dictionary has a fundamental flaw where a sensitive read might not be picked up as it is not defined in the dictionary. This sensitive read will then be passed as non-sensitive (even though the dictionary can be updated with these entries afterwards). Moreover, the sequences read are not always 100% accurate. Hence, sensitive reads might not be picked up even if the DNA segment is part of the dictionary due to sequence read errors. However, the bloom filter method has a built-in tolerance for reading errors.

All but one of the above solutions do not discuss their approach if the login credentials of research lab personal have been compromised or even if the data has been accessed by honest but curious research lab personnel.

4.2. DNA alignment

DNA read alignment (a process of aligning the read DNA strands to a reference genome) is the next significant step in genomic data preparation. DNA alignment is computationally intensive; hence, many research groups outsource this to a third party such as a Cloud provider [43]. A public Cloud provider is available for use by everyone, increasing the risk of data disclosure [14]. Also, Cloud service providers do not guarantee that an intruder cannot access the data [38].

4.2.1. Current solutions

Many solutions have been devised to address the safety of outsourcing the computation to an untrusted third party. Many security solutions rely on homomorphic encryption or one of its variants as a measure for protection. Using homomorphic encryption can take up to 5 minutes on 25 base pairs sequenced. An alternative privacy-preserving solution that utilises multiparty computation can take 4 seconds for 100 base pairs. These two approaches do not scale to a whole-genome sequence dataset containing multi-million base pairs [44].

Another option that is becoming more accessible is using a hybrid cloud. The speed on the hybrid cloud has improved by utilising a secure Seed-and-extend read mapping algorithm. The algorithm splits the computation such that the public cloud finds the exact seed matches using encrypted seeds, and the private cloud extends the seed matches using unencrypted data [44]. The second approach suggested by Popic et al. [45] is to preserve the read mapping's privacy for a hybrid cloud using BALAUR. BALAUR preserves read mapper for hybrid cloud based on locality-sensitive hashing and k-mer voting. It divides the computation between the trusted private client and the untrusted public cloud. It operates in two phases; the first phase identifies a few candidates' positions in the DNA strands where they can be aligned. These candidates are then assessed securely in the public cloud against an already hashed and indexed dictionary that was pre-prepared using the private client. This method is significantly faster than modern long read mappers, as the technique offloads 50–70% portion of the alignment to the cloud. However, Zhao et al. [46] created a new algorithm for aligning short reads where encrypted data is aligned in the public cloud while encryption and decryption occur in the private cloud. This algorithm produced results matching non-secure read mapping.

Another suggested method is the use of Intel's Software Guard Extension (SGX). This extension allows the user to create a protected enclave in an untrusted and less secure area [47]. SGX enclave has limited memory space, making it impractical for a large data set [48]. Sketching algorithms can be used alongside to address the memory limitation of Intel SGK. The sketching algorithm classifies and divides the original data, then re-structures it to fit into the Intel SGX enclave [48]. Lambert et al.

[49] introduced a novel method called MASK AI alongside Intel SGX. The utility provides a two-tier hybrid system; the first tier aligned masked reads in the public cloud while the second tier refines the first tier results.

Völp et al. [50] point out that adversaries can acquire variant information using the access patterns the algorithm generates despite using secure enclave alignment. The researchers present several solutions such as memory randomisation or cache access equalisation to hide access patterns or equalisation and keyed hashes, encrypting secret shuffle of the reference DNA.

4.2.2. Critical analysis

BALAUR uses a lot of memory and requires substantial network bandwidth [49]. While Zhao et al. [46] created an algorithm that works on short reads, most modern sequencers produce long reads which might render this approach to be less beneficial for real-world usage. The use of Intel SGX is limited by the size of the enclave which can vary depending on the number of processors and the memory size [51]. This approach could work for a small data set; however, adding processing power and memory can be proven to be costly if there is a need for a larger enclave in the cloud.

4.3. DNA data storage

Storing genomic data is the most common step after DNA alignment. However, there are no standard rules to imply the retention and return policies and where to store the data which means that research labs are expected to have their own standards. Most research labs store the genomic data in the patients' medical records. Doing this may result in unintentional or malicious access by a third party [52]. Vinatzer et al. [53] point out a lack of a mechanism to enforce adequate user authentication. Most databases do not implement strong password requirements by default, and access control is usually implemented when data is uploaded but rarely relevant when downloading digital DNA data. Elgabry et al. [14] highlighted that an adversary could gain access to genomic information by exploiting vulnerabilities within the database used to host the data. For example, they can exploit database authentication weakness in MongoDB (the database used by Genomics England) [54].

4.3.1. Current solutions

Secure storage for DNA and genomic data is vital to ensure data confidentiality, integrity, and authenticity. Huang et al. [55] introduced a novel method to reduce the storage requirement for alignment data called Selective retrieval on Encrypted and Compressed Reference–oriented Alignment Map (SECRAM) to reduce storage requirements while allowing selective genomic data retrieval. The approach enables random querying of subregions from genomic files in an encrypted form and preserves privacy during the downstream processes such as variant calling. Hwang et al. [56] presented an alternative solution to SECRAM to reduce the storage requirements for alignment data called Decentralised storage and compressed Reference-orientated alignment Map (D-RAM). The approach minimises the storage requirement by utilising reference base and bzip2 compression and preserves privacy by using the decentralised storage architecture.

Once the data is aligned, the outcome of the process can be a variety of genomic data. Homomorphic encryption can be used to encrypt stored genomic data; nevertheless, it is susceptible to brute force attacks [57]. Hosseini et al. [58] presented a tool to compress and encrypt FASTA files called CRYFA with low overhead DNA encryption and a compression capable of recognising various digital DNA file formats. CRYFA operates in two phases; phase one divides the DNA file into blocks and shuffles them, and phase two is to encrypt the file with AES standard encryption. CRYFA rearranges the file blocks to prevent an adversary from using low data complexity or Known-Plaintext-Attack (KPA) to decrypt the file.

Another encryption method that has been devised to overcome the possibilities of using a brute force attack against standard encryption methods is the use of honey encryption. Huang et al. [59] adapted honey encryption to encrypt genomic data. Genomic data files encrypted using honey encryption can be decrypted using any password entered; though, the correct genomic sequence will only appear if the correct password is used. In addition, this encryption method considers LD when encrypting genomic data. By considering genomic LD, this method avoids producing unrealistic genomic data when an adversary tries to access the encrypted data using a brute force attack.

Sousa et al. [60] discussed the rise of outsourcing storage to Cloud providers. They introduced a novel privacy-preserving algorithm to store a large amount of genomic data in a public Cloud. Their approach enables researchers to search for variants efficiently and in confidentiality while protecting data privacy. Their approach utilises optimal encoding for genomic data variants and combines it with homomorphic encryption and private information retrieval. Chen et al. [61] introduced a novel approach to storing genomic data in the cloud while balancing privacy and efficiency. The researchers utilised a graph-based database (Neo4j) with homomorphic encryption combined with Garbled Circuit.

4.3.2. Critical analysis

Using SECRAM to store alignment data seems a viable alternative to the de factor standards [43]. However, since the data is stored in centralised storage that the organisation manages, it might not be possible to guarantee the privacy of the data [56]. Storing data in a distributed storage environment when using D-RAM might not be feasible for some organisations due to cost or protecting their intellectual property. Using tools such as CRYFA to encrypt the stored data will protect the genomic data while at rest. However, it does not allow researchers to use the data while encrypted.

4.4. Summary

The DNA analysis and storage stage which includes sequencing pipeline and post-sequencing storage is at risk of unauthorised access and the disclosure of private information if the data is not adequately protected. Therefore, researchers have utilised various methods to prevent unauthorised data viewing. Figure 4 shows a summary of analysis and storing stage vulnerabilities. It consists of three levels; the top layer is the vulnerability source, the middle layer is the environment where the vulnerabilities can reside and the bottom layer is what can be done to mitigate or reduce the risk of these vulnerabilities.

Sensitive DNA sequence reads can be viewed if not sufficiently protected. Sequence privacy can be accomplished by using Classifications methods that classify reads into privacy-sensitive or non-

sensitive sections. Another approach is distributing the sequencing operation to multiple organisations, where each organisation will sequence a segment of the initial DNA.

Four methods can be used to protect DNA privacy during alignment i.e. encryption, classification, secure enclave and multiparty computation. Memory randomisation or cache equalisation can hide access patterns to the reference DNA while aligning using a hybrid cloud.

Encryption or encryption with compression and distributed storage can be used to preserve the privacy of the DNA data while stored (also known as data at rest) in a local, remote or cloud environment.

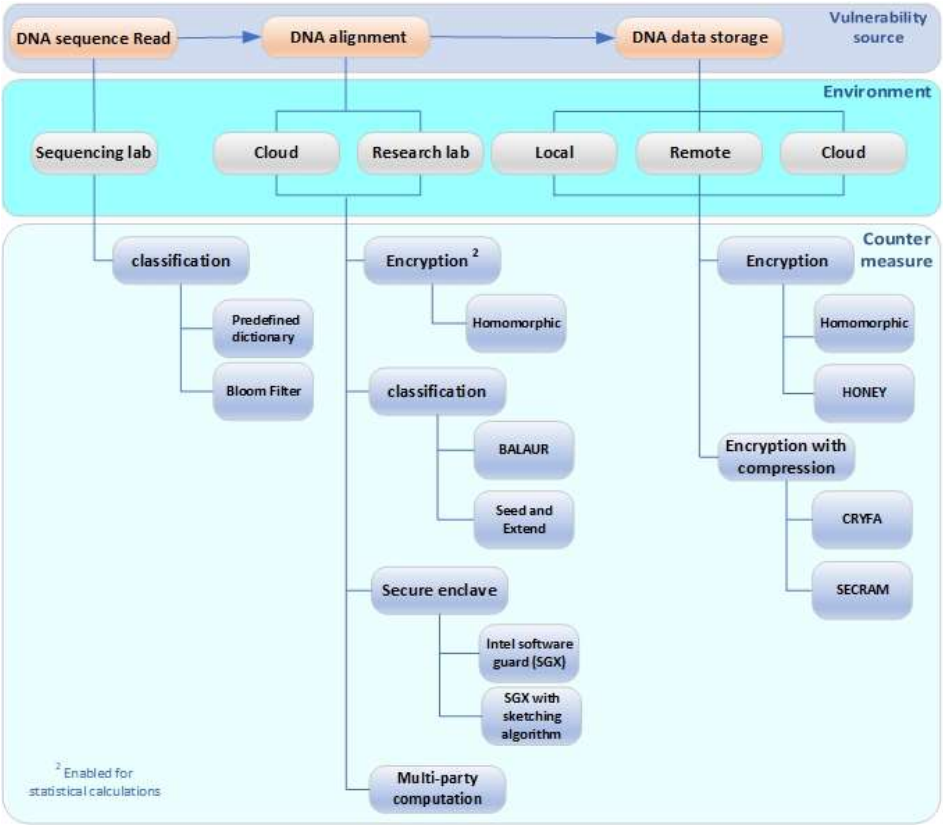

Figure 4. Summary of vulnerabilities associated with analysis and storing stage and their countermeasures

5. Querying, sharing, and Direct-to-Consumer stage vulnerabilities

5.1. Querying genomic data

Querying private genomic data is essential for personalised medicine, paternity, ancestry, and forensics. However, it constitutes a privacy risk to the participants' data.

5.1.1. Problem domain

According to Almadhoun et al. [62], membership inference attacks are the main vulnerability for genomic data owners. Samani et al. [63] showcased that correlation can be utilised for a genotype with hidden genomic data. Each individual has about 4 million differences in their genetic makeup to a reference sequence. It is possible to predict up to 40% of these differences with less than 1% error. This inference attack could happen if the adversary has access to genome data in the same population as the victim's data. This is achieved by relating genomic information to other publicly available information.

Henriksen-Bulmer & Jeary [64] highlighted aggregation of information method to identify individuals' genomic data. An adversary can identify an individual using aggregation by utilising multiple datasets, assuming that at least one of these data sets will include a social network or a search engine followed by a public dataset. An example of public datasets is shown in Table 2.

5.1.2. Current solutions

To reduce the risk of membership inference, Almadhoun et al. [62] stated that data owners attempt to reduce the risks by providing statistical answers to these queries. However, this approach has proven ineffective, as membership inference can be performed using the correlation between SNPs. To address this issue, Differential Privacy (DP) is used to protect the data. DP preserves privacy while sharing statistical information about a dataset by providing a mathematically rigorous approach (such as the Laplace mechanism) to prevent the risk of membership inference. The researchers debated the decreased effectiveness of DP when used on genomic data with interdependent data tuples (i.e. data structure that contains a number of elements) in the dataset.

Wang et al. [65] discussed the use of privacy-preserving computation for genomic data and showcased a novel method that utilises predicate encryption to query genomic data securely. The method is designed to help with precision medicine, where the patient genomic data is saved in the semi-trusted Cloud provider and accessed by a semi-trusted authorised party. The method has a low network overhead, but it is computationally intensive.

Ding et al. [66] suggested using a range query to query genomic data while maintaining privacy and security. The query is based on the Range proofs method which assures the requester that the required value is in the range provided. However, it does not disclose the actual value. Briguglio et al. [67] introduced a framework for ML with encryption that can predict a condition in a given genomic data while preserving its privacy. The researchers utilise ML predictive powers and homomorphic encryption to protect the privacy of the individuals in the genomic data set.

5.2. Sharing genomic data

Genomic research can provide a significant advantage in understanding health and disease, and it similarly presents promising prospects to speed up research by generating information-rich genome datasets. However, these benefits will only reach the production level if researchers and clinicians can access, compare and seek patterns in genomes belonging to many healthy and diseased individuals [68].

5.2.1. Data sharing limiting factors

Different data sources need to be brought together from multiple organisations to improve accuracy. As one organisation does not necessarily have all of the necessary information, several open-access genomic data sharing platforms appeared in the last decades; an example is shown in Table 2. However, sharing health data has to follow strict rules such as Health Insurance Portability and Accountability Act (HIPPA) in the USA. Also, organisations that attempt to share genomic data sources have the associated risk of privacy violation or informed consent violation and threat to participants' blood relatives [42].

Individual genomic data acts as a distinctive fingerprint that rarely changes; it includes sensitive information about the individual such as disease status or susceptibility to specific diseases. Sharing genomic information can also represent a privacy risk for family members as they correlate with the individual. An individual's genomic data can leak information about their family which can be accurately calculated through aggregate statistics. The process of predicting a family member's DNA can be achieved using the genetic dragnet method; this method is currently used for forensic

purposes by which DNA samples are gathered from the suspect's family to construct the suspect DNA [69]. Berger & Cho [70] demonstrated that the common practice of anonymising data to enable data sharing is ineffective against linkage attacks.

5.2.2. Sharing standards

There are mainly two systematic approaches to sharing genomic data. The first approach relies on having a central repository where all genomic data and associated information is kept. Genomics England uses this approach [71]. This approach allows researchers to log in and work on a unified dataset.

A second approach is a decentralised approach where each organisation keeps its data and allows access as a peer-to-peer network. For example, the BEACON project uses this approach [72].

BEACON [73] is a project by the Global Alliance For Genomic Health (GA4GH). Its purpose is to secure genomic data sharing. The BEACON project was designed to make it difficult for an adversary to re-identify an individual because the access is restrictive, and the researcher can only receive a "yes" or "no" to their genomic query [74].

Another approach for genomic data sharing which can be used as a centralised approach is Genome-Wide Association Study (GWAS) [75] or the decentralised approach which is the Federated GWAS [76]. GWAS is set up to provide a repository with a large population to produce reliable statistical results by using personal identifiable genetic markers. However, privacy concerns are making people reluctant to contribute [77]. For researchers, genomic data provide an immense benefit if combined with the patients' Electronic Health Records (EHRs). Hance, Harvard Medical School and the Massachusetts Institute of Technology (MIT) developed Informatics for Integrating Biology and the Bedside (i2b2) framework (now maintained by tranSMART Foundation). This solution can be implemented on a single site [78], [79] or can combine data from multiple sites [80].

Each of these approaches has its limitations. Storing data in a centralised location will act as a single point of failure. Another drawback is the reliance on the centralised location's ability to keep the

441 data private and confidential. A decentralised approach will require a higher cost to ensure data
442 security and privacy; also, it will require each site to maintain interoperable network security [81].

443 5.2.3. Current problems and solutions

444 There are flaws in how GWAS (whether centralised or federated) provides information e.g. Cai et al.
445 [82] presented a successful attack algorithm using genotype to identify individuals. He et al. [83]
446 demonstrated the ability to infer genotypes and phenotypes using genomic information of
447 individuals or the individuals' relatives information from GWAS based on belief propagation inserted
448 into a factor graph. Wang et al. [84] successfully evaluated two attacks types: trait inference and
449 identity inference based on Bayesian network through minging public GAWS statistics. Zhang et al.
450 [85] explained how exploiting GWAS statistics can infer traits from a given SNP genotype or a
451 genotype from a given trait or a trait from a given unknown trait. The researchers were able to infer
452 the information using three layers Bayesian network based on the Independence of Casual
453 Influences (ICI) modules.
454 To tackle some of the flows in GWAS, many researchers introduced novel methods to protect
455 participants' privacy. For instance, Zhang et al. [86] utilised secret sharing for multiparty
456 computation while utilising Hamming distance for secure sequence comparison. Wan et al. [87]
457 discussed sharing statistically aggregated genomic data (a statistically aggregated method for
458 anonymising genomic data began in the mid-2000s). This approach was aimed to standardise the
459 way genomic data is accessed through a centralised repository. While Bonte et al. [77] provided a
460 solution by combining homomorphic encryption with multiparty computation to provide accurate
461 statistics while preserving privacy. Privacy is achieved by returning yes/no to indicate a significant
462 correlation without revealing the statistical value itself.
463 Wu et al. [76] introduced a privacy-preserving framework for federated GWAS where genomic data
464 is computed locally within each participating institute, and only aggregated local statistics are
465 exchanged within the study network. Pascoal et al. [88] introduced Dynamic, Private and Secure
466 (DyPS) GWAS which is a federated GWAS system where each biocentre shares its statistics without

revealing its data. All statistics are computed securely within Intel SGX while preserving privacy by safely releasing aggregated statistics after passing several privacy checks i.e. Likelihood-ratio test. Wang et al. [89] pointed out that the current GWAS privacy-preserving solutions focused on protecting individuals. If an attacker compromised GAWS statistics and identified an individual, the attacker could infer information regarding the individual's relatives using the Transmission Disequilibrium Test (TDT). The researchers developed a privacy solution to protect the families' privacy built on differential privacy using the Shortest Hamming Distance (SHD) score method which balanced privacy and utility.

With all the suggested modifications of the GAWS results, Halimi et al. [25] pointed out that the researchers must verify the accuracy of the results obtained from the GWAS, especially if the results source data have noise to maintain differential privacy. The authors devised a framework for result verification while preserving the data's privacy; they achieved this by probabilistically calculating the correctness of the results.

Simultaneously, other researchers showed some of the drawbacks of the BEACON platform. Even with such restrictions, it is possible to identify individuals with an accuracy of 95% by using the Likelihood Ratio Test (LRT) [90].

Raisaro et al. [91] proposed three approaches to reduce the risk of re-identification in BEACON. The first approach costs the number of accesses per user for each genome, while the other two manipulate the system to obfuscate the presence of the rare alleles. Demmler et al. [92] provided a solution that can be an addon to secure BEACON. The researchers' solution allows private multi-variant and multi-property queries that obfuscate which elements it accessed and what parts match the query to private aggregated data from multiple sources.

Raisaro et al. [78] pointed out that i2b2 cohort explorer lacked protection beyond patient de-identification and access control and presented a privacy-preserving solution based on encrypting

patients' data with *somewhat* homomorphic encryption and delivering the results with the concept of differential privacy.

Human genomic data sharing plays a big part in understanding health and disease as a result. Many researchers try to introduce new approaches to preserve participants privacy while using and sharing the data. For example, Chen et al. [93] presented PRINCESS, a framework for international collaboration to analyse rare disease genetic data while safeguarding patients' privacy. PRINCESS utilise SGX to facilitate secure and distributed computations. Raisaro et al. [81] suggested using homomorphic encryption and its variants to secure shared genomic data. It allows other parties to query the data while the data is encrypted. The researchers introduced a new approach for sharing genomic information via MedCo which is a system that allows many organisations and clinics to share their data in a hybrid decentralised system by distributing trusts between the storage and processing units to form a federated incorporated network. Schneider et al. [94] designed an efficient distributed privacy-preserving protocol that is based on multiparty computation using approximated Edit Distance(ED) to protect Similar Sequence Queries (SSQs). A new method has been suggested by Ozercan et al. [95] for multiparty data sharing which uses blockchain; the method uses a decentralised approach in storing the data. The blockchain method integrates with the existing solutions used in different organisations. Another approach for using blockchain was introduced by Grishin et al. [96] where genomic data is encrypted and shared by multiple independent parties. The encryption key is split between parties. Any request to access the data and user consent is stored in a blockchain.

Some researchers direct their efforts to secure specific fields in genomic studies e.g Gürsoy et al. [97] introduced a new method to reduce private information leakage from functional genomics. The researchers presented techniques to minimise common privacy risks that were quantified by adopting statistical techniques. Jagadeesh et al. [98] provided a secure multiparty computation for genomic diagnoses without revealing patient genomes based on two approaches. The first approach transforms the patient genome into vectors that indicate the relevant variants after simple

519 operations. The second approach uses a cryptographic method to perform private computations.
520 Akgün et al. [99] produced a privacy-preserving multiparty computation approach to identify
521 disease-associated variants and genes based on a combination of arithmetic and boolean sharing in
522 the same computation. The researchers' approach was faster and more accurate than the previous
523 solution, and It could also allow cross-institution collaborations which were very useful in the case of
524 rare diseases.

525 Sharing genomic information securely is important irrespective of which approach or tool an
526 organisation uses to share their data. An adversary can sniff data packets that research institutes
527 send and receive to obtain sensitive genomic data such as using shodan.io, a data-sharing tool used
528 by many research institutes [8]. It is essential to protect the network traffic. Kelleher et al. [100]
529 created a protocol to obtain shared genomic data called htsget which is based on HTTP(s) GET
530 requests and it works with Transport Layer Security (TLS) encryption which uses OAuth2.0 tokens to
531 authorise data requests.

532 Wan et al. [87] point out the need for a trade-off between privacy and utility; they highlighted that
533 concentrating on what is possible might not be probable; the researchers use Game theory to
534 provide a method to measure risk vs protection. This can help data sharers to find the best
535 protection strategy.

536

537

538 *Table 2: An example of Genomic Public data sets (source genomic data sets websites)*

Genomics Public data set	URL
OpenSNP	https://opensnp.org/
Genome-Wide Association Studies (GWAS)	https://www.ebi.ac.uk/gwas/
Global Alliance for Genomics and Health (GA4GH)	https://www.ga4gh.org/
1000 Genomes Project	https://www.internationalgenome.org/data/
The 100,000 Genomes Project	https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/
The Cancer Genome Atlas (TCGA)	https://www.sevenbridges.com/tcga/
International Cancer Genome Consortium	https://daco.icgc.org/

Genome in a Bottle	https://jimb.stanford.edu/giab-resources
National Institutes of Health (NIH)	https://www.nih.gov/
The Human Connectome Project	http://www.humanconnectomeproject.org/
Pan-UK Biobank	https://pan.ukbb.broadinstitute.org/
Nucleotide BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
RefGenie	http://refgenie.databio.org/en/latest/
GnomAD	https://gnomad.broadinstitute.org/
Open Targets	https://www.opentargets.org/

5.1. Direct-to-Consumer genetic testing

DTC genetic testing is another threat to privacy. These companies collect genomic data from individuals who may not fully understand the full impact on themselves or their families and future blood relatives. Some DTC companies and the services they provide are listed in Table 3. DTC uses the genomic data beyond the service provided, as the terms of the service for most of them do not clearly state how customers' data will be used or whom the data will be shared with. DTC privacy threats stem from the fact that they are not a health provider. Hence, they do not have to follow the same rules and regulations imposed on health care providers such as HIPPA in the USA [101]. Laestadius et al. [102] found that DTC does not provide sufficient information regarding how their data will be treated. They also found that most DTC companies fail to mention the risks of re-identification and genetic discrimination.

DNA and genomic data production are very beneficial for genuine research and usage purposes. Nevertheless, genomic data are similarly commercially very valuable. For example, in 2018, GlaxoSmith Kline bought thousands of customers personal data from a commercial DNA testing kit provider, 23AndMe, for $300 million [103].

Table 3: Popular Direct-to-Consumers (DTC) companies, the approximate customer numbers and the primary service provided by them (source DTC websites)

Direct-to-Consumer Company	Consumer Numbers	Service Provided	Notes
GedMatch (https://www.gedmatch.com)	1.3 Million	Autosomal DNA genealogy service	Owned by Verogen (forensic science & sequencing), the GedMatch database was breached by hackers in July 2020
Ancestry (https://www.ancestry.co.uk)	over 15 Million	Autosomal DNA genealogy and family history service	
23andMe (https://www.23andme.com)	12 Million	Autosomal DNA genealogy and health predisposition service	

My heritage (https://www.myheritage.com)	4.65 Million	Autosomal DNA genealogy and family history service	
FTDNA (https://www.familytreedna.com)	951 thousand	Autosomal DNA and mitochondria DNA genealogy service	
Genome link (https://genomelink.io)	No data	Genetic trait analysis service	Users don't need to keep their data on the site
I search me (https://www.ichrogene.com)	No Data	Genetic trait analysis service	

Sharing genomic data via DTC websites as shown in Table 3 or via clinical research websites as shown in Table 3, has its own associated risk of re-identification. Bonomi et al. [101] showcased various methods such as anonymising genomic data with health privacy to reduce the risk of re-identification. Health privacy is a method that masks SNPs and limits the disclosure of sensitive phenotypes of the genomic data. The authors also highlighted the recent development in regulations and guidelines to preserve consumers' privacy in a DTC setting even though it is in its early stages.

Ney et al. [104] examined the open design and the broad Application programming interface (API) offered by some DTC websites. The researchers showcased the number of security vulnerabilities in GEDmatch API and demonstrated the ability of an adversary to extract a large percentage of the genetics markers from other users (including medically sensitive markers) by typically formatting genetic data files and running standard queries.

Voluntary best practices for genetic information use and security are being established by The Future of Privacy Forum (FPF) [105] which is working with leading DTC companies (23andMe, Ancestry, Helix, MyHeritage, and Habit) and promotes transparency in the way that the data is used. The Future of Privacy Forum is also working on enhanced consumer protection and consumer consent to encourage people to donate their DNA for research [106].

Hansson et al. [107] questioned the need to change the regulatory requirements in order to increase the protection for genomic data; the researchers pointed out that stricter legal regulations will be detrimental to genomic research. The researchers discussed the term "harm" caused by leaked genomic data to the study participants and the need to balance it with the benefits, especially when it comes to rare genetic disorders.

5.2. Summary

DNA data is at risk of re-identification attacks when the data is queried and shared. There is also the vulnerability associated with how genomic information is shared and used in DTC settings. Figure 5 shows a summary of the risks and their countermeasures. It consists of three levels; the top section is the vulnerability source, the middle section is the attack vector and the bottom section demonstrates the methods used to mitigate or reduce the risk of that attack vector.

Genomic data querying vulnerability sources have three associated attack vectors: aggregation of information, aggregation of statistics and correlation attacks. To reduce the risk of these threats, differential privacy, range query, encryption with statistical calculation capability, privacy-preserving computation or ML with encryption methods can be used as countermeasures.

Sharing genomic data have many privacy risks i.e. belief propagation, inference, linkage and likelihood ratio test attacks. However, several countermeasures can be utilised to secure data sharing and preserve data privacy such as sharing statistical results, statistically aggregated data, using multiparty data sharing, and multiparty computation with secret sharing and many others; there is also the need to use Transport Layer Security (TLS) when sending and receiving shared genomic data.

DTC emerged as a significant threat to genomic privacy as it is not always clear how customers' data will be used due to the complexity of the DTCs' terms and conditions. There are also many vulnerabilities associated with the DTCs' websites such as the ability to identify individuals through carefully constructed queries, coupled with vulnerabilities with the DCTs' websites API. To countermeasure these attack vectors, DTCs should use best practice guidelines introduced by the future of privacy forums coupled with anonymising genomic data using health privacy.

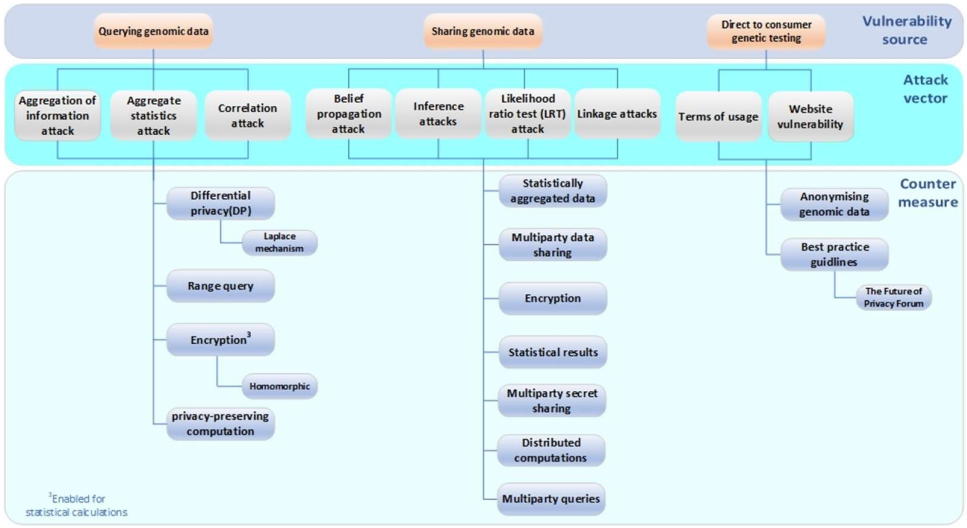

Figure 5. Summary of vulnerabilities associated with querying, sharing and direct to consumer genomic testing stage and their countermeasures

6. Conclusion

Genomic research is vital in finding new treatments and understanding complex diseases, plays an essential role in forensics and understanding our heritage. Equally, genomic security is fundamental to one's privacy. There are many attempts to secure genomic data; however, some of these solutions fall short in protecting our genomic data or do not scale to cover actual life data.

In this overview, the term digital DNA life cycle has been introduced, digital DNA data privacy, security threats and possible countermeasures have been investigated. The overview covers the threats to pre-digital DNA and throughout the digital DNA lifecycle and shows that the DNA is under threat at every stage. At the pre-digital DNA stage, DNA that is obtained from non trusted sources can disrupt the sequencing cycle or create a worm that can infect the downstream computers or trojans that can be used to target DNA sequencer hardware hance, DNA source authenticity and security is paramount. There are also vulnerabilities in DNA sequencing software where insecure function calls within the software can cause side-channel attacks or allow the attacker arbitrary code execution. This can be avoided by following software development security best practices.

There are many privacy risks throughout the digital DNA lifecycle such as threats stemming from DNA sequence reads, sequence alignments and storage where data can be viewed if not sufficiently

protected, or the danger of individuals being identified by an attacker while Querying genomic data using various methods such as aggregation of information attack or correlation attacks same goes for genomic data sharing where linkage and likelihood ratio test attacks can be used to identify participants. Some of the methods used to manage the risks are differential privacy, data aggregation and encryption. Another threat to privacy has risen from DTC as an attacker can identify individuals through carefully constructed queries, coupled with vulnerabilities with the DCTs' websites API. DTC companies should utilise best practice guidelines while anonymising their health data using health privacy to reduce their customers' risks.

Real-time checking, combining adaptive security solutions, e.g. the use of ML to detect illegitimate access coupled with developing international regulations and awareness of these risks, etc., would increase confidence in genomic privacy and encourage more donors to participate in research. However, there is also a need for these security and privacy solutions not to slow down or add extra burdens on the researchers to take full advantage of what genomic research can provide.

7. References

[1] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux, "De-anonymizing Genomic Databases Using Phenotypic Traits," in *Proceedings on Privacy Enhancing Technologies*, 2015, vol. 2.

[2] M. Backes, P. Berrang, M. Humbert, X. Shen, and V. Wolf, "Simulating the Large-Scale Erosion of Genomic Privacy over Time," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 5, pp. 1405–1412, 2018.

[3] D. Sero *et al.*, "Facial recognition from DNA using face-to-DNA classifiers," *Nat. Commun.*, vol. 10, no. 1, pp. 1–12, Dec. 2019.

[4] C. Lippert *et al.*, "Identification of individuals by trait prediction using whole-genome sequencing data.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 38, pp. 10166–10171, 2017.

[5] L. Qiao *et al.*, "Genome-wide variants of Eurasian facial shape differentiation and a prospective model of DNA based face prediction," *J. Genet. Genomics*, vol. 45, no. 8, pp. 419–432, 2018.

[6] S. Richmond, L. J. Howe, S. Lewis, E. Stergiakouli, and A. Zhurov, "Facial Genetics: A Brief Overview," *Front. Genet.*, vol. 9, p. 462, 2018.

[7] M. Elgabry, D. Nesbeth, and S. D. Johnson, "A Systematic Review of the Criminogenic Potential of Synthetic Biology and Routes to Future Crime Prevention," *Front. Bioeng. Biotechnol.*, vol. 8, p. 1119, 2020.

[8] Y. Joly, I. N. Feze, L. Song, and B. M. Knoppers, "Comparative Approaches to Genetic

Discrimination: Chasing Shadows?," *Trends Genet.*, vol. 33, no. 5, pp. 299–302, 2017.

[9] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Quantifying Interdependent Risks in Genomic Privacy," *ACM Trans. Priv. Secur.*, vol. 20, no. 1, pp. 1–31, Feb. 2017.

[10] E. Zeggini, A. L. Gloyn, A. C. Barton, and L. V Wain, "Translational genomics and precision medicine: Moving from the lab to the clinic," *Science (80-.).*, vol. 365, no. 6460, pp. 1409 LP – 1413, Sep. 2019.

[11] Y. Y. Liu and S. A. Harbison, "A review of bioinformatic methods for forensic DNA analyses," *Forensic Science International: Genetics*, vol. 33. Elsevier Ireland Ltd, pp. 117–128, 01-Mar-2018.

[12] N. Moray, K. E. Pink, P. Borry, and M. H. D. Larmuseau, "Paternity testing under the cloak of recreational genetics," *Eur. J. Hum. Genet.*, vol. 25, no. 6, pp. 768–770, Mar. 2017.

[13] L. Vossaert, I. Chakchouk, R. Zemet, and I. B. Van den Veyver, "Overview and recent developments in cell-based noninvasive prenatal testing," *Prenat. Diagn.*, vol. 41, no. 10, pp. 1202–1214, May 2021.

[14] M. Akgün, A. O. Bayrak, B. Ozer, and M. Ş. Sağiroğlu, "Privacy preserving processing of genomic data: A survey," *J. Biomed. Inform.*, vol. 56, pp. 103–111, 2015.

[15] D. Lu *et al.*, "Methods of privacy-preserving genomic sequencing data alignments," *Brief. Bioinform.*, May 2021.

[16] M. M. Al Aziz *et al.*, "Privacy-preserving techniques of genomic data-a survey," *Brief. Bioinform.*, vol. 20, no. 3, pp. 887–895, 2017.

[17] A. Mittos, B. Malin, and E. De Cristofaro, "Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective," *Proc. Priv. Enhancing Technol.*, no. 1, pp. 87–107, 2019.

[18] X. Shi and X. Wu, "An overview of human genetic privacy," *Ann. N. Y. Acad. Sci.*, vol. 1387, no. 1, pp. 61–72, 2017.

[19] A. P. Schwab, H. S. Luu, J. Wang, and J. Y. Park, "Genomic Privacy," *Clin. Chem.*, vol. 64, no. 12, pp. 1696–1703, 2018.

[20] A. B. Carter, "Considerations for Genomic Data Privacy and Security when Working in the Cloud," *J. Mol. Diagnostics*, vol. 21, no. 4, pp. 542–552, 2019.

[21] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6. Nature Publishing Group, pp. 409–421, 2014.

[22] A. Mohammed Yakubu and Y.-P. P. P. P. Chen, "Ensuring privacy and security of genomic data and functionalities," *Brief. Bioinform.*, vol. 21, no. 2, pp. 511–526, Mar. 2020.

[23] B. Abinaya and S. Santhi, "A survey on genomic data by privacy-preserving techniques perspective," *Comput. Biol. Chem.*, vol. 93, p. 107538, Aug. 2021.

[24] M. Naveed *et al.*, "Privacy in the Genomic Era," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–44, 2015.

[25] A. Halimi *et al.*, "Privacy-Preserving and Efficient Verification of the Outcome in Genome-Wide Association Studies," *ArXiv*, vol. abs/2101.0, 2021.

[26] L. Hartwell, M. Goldberg, J. Fischer, and L. Hood, *Genetics: from genes to genomes*. McGraw-Hill Education, 2018.

[27] K. L. Howe *et al.*, "Ensembl 2021," *Nucleic Acids Res.*, vol. 49, no. 2, 2021.

[28] V. Jalili, M. Matteucci, M. Masseroli, and S. Ceri, “Indexing Next-Generation Sequencing data,” *Inf. Sci. (Ny).*, vol. 384, pp. 90–109, Apr. 2017.

[29] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, “Demonstration of end-to-end Automation of DNA Data storage,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–5, 2019.

[30] P. Ney, K. Koscher, L. Organick, L. Ceze, and T. Kohno, “Computer security, privacy, and DNA sequencing: compromising computers with synthesized DNA, privacy leaks, and more,” in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 765–779.

[31] J. E. Gallegos, D. M. Kar, I. Ray, I. Ray, and J. Peccoud, “Securing the Exchange of Synthetic Genetic Constructs Using Digital Signatures,” *ACS Synth. Biol.*, vol. 9, no. 10, pp. 2656–2664, Oct. 2020.

[32] S. Allen Morse, G. D. Koblentz, J. Diggans, and E. Leproust, “POLICY AND PRACTICE REVIEWS Next Steps for Access to Safe, Secure DNA Synthesis,” *Front. Bioeng. Biotechnol.*, vol. 1, p. 86, 2019.

[33] S. S. Ali, M. Ibrahim, J. Rajendran, O. Sinanoglu, and K. Chakrabarty, “Supply-Chain Security of Digital Microfluidic Biochips,” *Computer (Long. Beach. Calif).*, vol. 49, no. 8, pp. 36–43, Aug. 2016.

[34] I. Fayans, Y. Motro, L. Rokach, Y. Oren, and J. Moran-Gilad, “Cyber security threats in the microbial genomics era: implications for public health,” *Eurosurveillance*, vol. 25, no. 6, p. 1900574, Feb. 2020.

[35] Q. Li *et al.*, “Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform,” *BMC Genomics*, vol. 20, no. 1, p. 215, 2019.

[36] A. Mitra, M. Skrzypczak, K. Ginalski, and M. Rowicka, “Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform,” *PLoS One*, vol. 10, no. 4, p. 120520, Apr. 2015.

[37] S. Arshad, J. Arshad, M. M. Khan, and S. Parkinson, “Analysis of security and privacy challenges for DNA-genomics applications and databases,” *J. Biomed. Inform.*, vol. 119, p. 103815, Jul. 2021.

[38] J. Decouchant, M. Fernandes, M. Völp, F. M. Couto, and P. Esteves-Veríssimo, “Accurate filtering of privacy-sensitive information in raw genomic data,” *J. Biomed. Inform.*, vol. 82, pp. 1–12, 2018.

[39] V. V. Cogo, A. Bessani, F. M. Couto, and P. Verissimo, “A High-Throughput Method to Detect Privacy-Sensitive Human Genomic Data,” in *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society - WPES ’15*, 2015, pp. 101–110.

[40] M. Fernandes, J. Decouchant, M. Volp, F. M. Couto, and P. Esteves-Verissimo, “DNA-SeAl: Sensitivity Levels to Optimize the Performance of Privacy-Preserving DNA Alignment,” *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 3, pp. 907–915, Mar. 2020.

[41] A. Gholami, M. A. Maddah-Ali, and S. Abolfazl Motahari, “Private Shotgun DNA Sequencing,” 2019, vol. 2019-July, pp. 171–175.

[42] M. Z. Hasan, M. S. R. Mahdi, M. N. Sadat, and N. Mohammed, “Secure count query on encrypted genomic data,” *J. Biomed. Inform.*, vol. 81, pp. 41–52, 2018.

[43] S. Roy *et al.*, “Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists,” *J. Mol. Diagnostics*, vol. 20, no. 1, pp. 4–27, 2018.

[44] V. Popic and S. Batzoglou, “Privacy-Preserving Read Mapping Using Locality Sensitive Hashing

and Secure Kmer Voting," *bioRxiv*, p. 46920, Jan. 2016.

[45] V. Popic and S. Batzoglou, "A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy," *Nat. Commun.*, vol. 8, no. 1, p. 15311, May 2017.

[46] Y. Zhao, X. Wang, and H. Tang, "A Secure Alignment Algorithm for Mapping Short Reads to Human Genome," *J. Comput. Biol.*, vol. 25, no. 6, pp. 529–540, May 2018.

[47] F. Chen *et al.*, "PRESAGE: PRivacy-preserving gEnetic testing via SoftwAre Guard Extension," *BMC Med. Genomics*, vol. 10, no. S2, p. 48, 2017.

[48] C. Kockan *et al.*, "Sketching algorithms for genomic data analysis and querying in a secure enclave," *Nat. Methods*, vol. 17, no. 3, pp. 295–301, 2020.

[49] C. Lambert, M. Fernandes, J. Decouchant, and P. Esteves-Verissimo, "MaskAl: Privacy Preserving Masked Reads Alignment using Intel SGX," in *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*, 2018, pp. 113–122.

[50] M. Völp, J. Decouchant, C. Lambert, M. Fernandes, and P. Esteves-Verissimo, "Enclave-Based Privacy-Preserving Alignment of Raw Genomic Information," in *Proceedings of the 2nd Workshop on System Software for Trusted Execution - SysTEX'17*, 2017, pp. 1–6.

[51] "Intel® Software Guard Extensions (Intel® SGX)." [Online]. Available: https://www.intel.co.uk/content/www/uk/en/architecture-and-technology/software-guard-extensions.html. [Accessed: 19-Oct-2021].

[52] M. Shabani, D. Vears, and P. Borry, "Raw Genomic Data: Storage, Access, and Sharing," *Trends Genet.*, vol. 34, no. 1, pp. 8–10, Jan. 2018.

[53] B. A. Vinatzer, L. S. Heath, H. M. J. Almohri, M. J. Stulberg, C. Lowe, and S. Li, "Cyberbiosecurity Challenges of Pathogen Genome Databases," *Front. Bioeng. Biotechnol.*, vol. 7, p. 106, 2019.

[54] Genomics England, "Genomics England," 2021. [Online]. Available: https://www.genomicsengland.co.uk/. [Accessed: 06-Feb-2021].

[55] Z. Huang *et al.*, "A privacy-preserving solution for compressed storage and selective retrieval of genomic data.," *Genome Res.*, vol. 26, no. 12, pp. 1687–1696, 2016.

[56] D. Hwang, S. Choi, J. Shin, G. Song, and Y. Choi, "Privacy-Preserving Compressed Reference-Oriented Alignment Map Using Decentralized Storage," *IEEE Access*, vol. 6, pp. 45990–46001, 2018.

[57] J.-P. Aumasson, "The impact of quantum computing on cryptography," *Comput. Fraud Secur.*, vol. 2017, no. 6, pp. 8–11, 2017.

[58] M. Hosseini, D. Pratas, and A. J. Pinho, "Cryfa: a secure encryption tool for genomic data," *Bioinformatics*, vol. 35, no. 1, pp. 146–148, 2019.

[59] Z. Huang, E. Ayday, J. Fellay, J.-P. Hubaux, and A. Juels, "GenoGuard: Protecting Genomic Data against Brute-Force Attacks," in *2015 IEEE Symposium on Security and Privacy*, 2015, pp. 447–462.

[60] J. S. Sousa *et al.*, "Efficient and secure outsourcing of genomic data storage," *BMC Med. Genomics*, vol. 10, no. 2, pp. 15–28, 2017.

[61] L. Chen, M. M. Aziz, N. Mohammed, and X. Jiang, "Secure large-scale genome data storage and query," *Comput. Methods Programs Biomed.*, vol. 165, pp. 129–137, Oct. 2018.

[62] N. Almadhoun, E. Ayday, and Ö. Ulusoy, "Inference attacks against differentially private query

results from genomic datasets including dependent tuples," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i136–i145, Jul. 2020.

[63] S. S. Samani *et al.*, "Quantifying Genomic Privacy via Inference Attack with High-Order SNV Correlations," in *2015 IEEE Security and Privacy Workshops*, 2015, pp. 32–40.

[64] J. Henriksen-Bulmer and S. Jeary, "Re-identification attacks—A systematic literature review," *Int. J. Inf. Manage.*, vol. 36, no. 6, Part B, pp. 1184–1192, 2016.

[65] B. Wang, W. Song, W. Lou, and Y. T. Hou, "Privacy-preserving pattern matching over encrypted genetic data in cloud computing," in *Proceedings - IEEE INFOCOM*, 2017, pp. 1–9.

[66] X. Ding, E. Ozturk, and G. Tsudik, "Balancing security and privacy in genomic range queries," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2019, pp. 106–110.

[67] W. Briguglio, P. Moghaddam, W. A. Yousef, I. Traoré, and M. Mamun, "Machine learning in precision medicine to preserve privacy via encryption," *Pattern Recognit. Lett.*, vol. 151, pp. 148–154, 2021.

[68] T. Haeusermann, B. Greshake, A. Blasimme, D. Irdam, M. Richards, and E. Vayena, "Open sharing of genomic data: Who does it and why?," *PLoS One*, vol. 12, no. 5, p. e0177158, 2017.

[69] E. Ayday and M. Humbert, "Inference Attacks against Kin Genomic Privacy," *IEEE Secur. Priv.*, vol. 15, no. 5, pp. 29–37, 2017.

[70] B. Berger and H. Cho, "Emerging technologies towards enhancing privacy in genomic data sharing," *Genome Biol.*, vol. 20, no. 1, p. 128, 2019.

[71] Genomics England, "Genome Sequencing," 2021. [Online]. Available: https://www.genomicsengland.co.uk/understanding-genomics/genome-sequencing/. [Accessed: 25-Apr-2021].

[72] A. Page *et al.*, "A federated ecosystem for sharing genomic, clinical data," *Am. Assoc. Adv. Sci.*, vol. 352, no. 6291, pp. 1278 LP – 1280, Jun. 2016.

[73] GA4GH, "BEACON," 2021. [Online]. Available: https://beacon-project.io/. [Accessed: 14-Aug-2020].

[74] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, "Aftermath of bustamante attack on genomic beacon service," *BMC Med. Genomics*, vol. 10, no. S2, p. 43, 2017.

[75] T. Beck, T. Shorter, and A. J. Brookes, "GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D933–D940, Jan. 2020.

[76] X. Wu *et al.*, "A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis," *Brief. Bioinform.*, vol. 22, no. 3, May 2021.

[77] C. Bonte, E. Makri, A. Ardeshirdavani, J. Simm, Y. Moreau, and F. Vercauteren, "Towards practical privacy-preserving genome-wide association study," *BMC Bioinformatics*, vol. 19, no. 1, p. 537, 2018.

[78] J. L. Raisaro *et al.*, "Protecting Privacy and Security of Genomic Data in i2b2 With Homomorphic Encryption and Differential Privacy," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–1, 2018.

[79] TranSMART, "I2b2," 2021. [Online]. Available: https://www.i2b2.org/. [Accessed: 17-Jul-2021].

[80] J. G. Klann, A. Abend, V. A. Raghavan, K. D. Mandl, and S. N. Murphy, “Data interchange using i2b2,” *J. Am. Med. Inform. Assoc.*, vol. 23, no. 5, pp. 909–915, 2016.

[81] J. L. Raisaro *et al.*, “MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 16, no. 4, pp. 1328–1341, 2018.

[82] R. Cai *et al.*, “Deterministic identification of specific individuals from GWAS results,” *Bioinformatics*, vol. 31, no. 11, pp. 1701–1707, Jun. 2015.

[83] Z. He, J. Yu, J. Li, Q. Han, G. Luo, and Y. Li, “Inference Attacks and Controls on Genotypes and Phenotypes for Individual Genomic Data,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 17, no. 3, pp. 1–1, May 2018.

[84] Y. Wang, J. Wen, X. Wu, and X. Shi, “Infringement of Individual Privacy via Mining Differentially Private GWAS Statistics,” in *International Conference on Big Data Computing and Communications*, 2016, pp. 355–366.

[85] L. Zhang, Q. Pan, Y. Wang, X. Wu, and X. Shi, “Bayesian network construction and genotype-phenotype inference using GWAS statistics,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 16, no. 2, pp. 475–489, Mar. 2019.

[86] Y. Zhang, M. Blanton, and G. Almashaqbeh, “Secure distributed genome analysis for GWAS and sequence comparison computation,” *BMC Med. Inform. Decis. Mak.*, vol. 15 Suppl 5, no. S5, pp. S4–S4, 2015.

[87] Z. Wan, Y. Vorobeychik, E. W. Clayton, M. Kantarcioglu, and B. Malin, “Game theory for privacy-preserving sharing of genomic data,” in *Responsible Genomic Data Sharing*, X. Jiang and H. B. T.-R. G. D. S. Tang, Eds. Academic Press, 2020, pp. 135–160.

[88] T. Pascoal, J. Decouchant, A. Boutet, and P. Esteves-Verissimo, “DyPS: Dynamic, Private and Secure GWAS,” in *Proceedings on Privacy Enhancing Technologies*, 2021, vol. 2021, no. 2, pp. 214–234.

[89] M. Wang *et al.*, “Mechanisms to protect the privacy of families when using the transmission disequilibrium test in genome-wide association studies,” *Bioinformatics*, vol. 33, no. 23, pp. 3716–3725, Dec. 2017.

[90] N. Von Thenen, E. Ayday, and A. E. Cicek, “Re-identification of individuals in genomic data-sharing beacons via allele inference,” *Bioinformatics*, vol. 35, no. 3, pp. 365–371, 2019.

[91] J. L. Raisaro *et al.*, “Addressing Beacon re-identification attacks: Quantification and mitigation of privacy risks,” *J. Am. Med. Informatics Assoc.*, vol. 24, no. 4, 2017.

[92] D. Demmler, K. Hamacher, T. Schneider, and S. Stammler, “Privacy-preserving whole-genome variant queries,” in *6th International Conference on Cryptology and Network Security,* 2018, vol. 11261 LNCS, pp. 71–92.

[93] F. Chen *et al.*, “PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS,” *Bioinformatics*, vol. 33, no. 6, pp. 871–878, Mar. 2017.

[94] T. Schneider and O. Tkachenko, “Towards Efficient Privacy-Preserving Similar Sequence Queries on Outsourced Genomic Databases,” in *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, 2018, pp. 71–75.

[95] H. I. Ozercan, A. M. Ileri, E. Ayday, and C. Alkan, “Realizing the potential of blockchain technologies in genomics.,” *Genome Res.*, vol. 28, no. 9, pp. 1255–1263, 2018.

[96] D. Grishin, K. Obbad, and G. M. Church, “Data privacy in the age of personal genomics,” *Nat.*

Biotechnol., vol. 37, no. 10, pp. 1115–1117, 2019.

[97] G. Gürsoy *et al.*, "Data Sanitization to Reduce Private Information Leakage from Functional Genomics," *Cell*, vol. 183, no. 4, pp. 905-917.e16, 2020.

[98] K. A. Jagadeesh *et al.*, "Deriving genomic diagnoses without revealing patient genomes," *Science (80-.).*, vol. 357, no. 6352, pp. 692–695, Aug. 2017.

[99] M. Akgün, A. B. Ünal, B. Ergüner, N. Pfeifer, and O. Kohlbacher, "Identifying disease-causing mutations with privacy protection," *Bioinformatics*, vol. 36, no. 21, pp. 5205–5213, Nov. 2020.

[100] J. Kelleher *et al.*, "Htsget: A protocol for securely streaming genomic data," *Bioinformatics*, vol. 35, no. 1, pp. 119–121, 2019.

[101] L. Bonomi, Y. Huang, and L. Ohno-Machado, "Privacy challenges and research opportunities for genomic data sharing," *Nat. Genet.*, vol. 52, pp. 646–654, 2020.

[102] L. I. Laestadius, J. R. Rich, and P. L. Auer, "All your data (effectively) belong to us: Data practices among direct-to-consumer genetic testing firms," *Genet. Med.*, vol. 19, pp. 513–520, 2017.

[103] L. Defrancesco and A. Klevecz, "Your DNA broker," *Nat. Biotechnol.*, vol. 37, no. 10, pp. 842–847, 2019.

[104] P. Ney, L. Ceze, and T. Kohno, "Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference," in *Network and Distributed System Security Symposium (NDSS)*, 2020.

[105] The Future of PrivacyForum, "The Future of Privacy Forum." [Online]. Available: https://fpf.org/. [Accessed: 20-Apr-2021].

[106] R. M. Hendricks-Sturrup and C. Y. Lu, "Direct-to-Consumer Genetic Testing Data Privacy: Key Concerns and Recommendations Based on Consumer Perspectives," *J. Pers. Med.*, vol. 9, no. 2, 2019.

[107] M. G. Hansson *et al.*, "The risk of re-identification versus the need to identify individuals in rare disease research," *Eur. J. Hum. Genet.*, vol. 24, no. 11, pp. 1553–1558, 2016.