



LJMU Research Online

Alsaffar, MM, Hasan, M, McStay, GP and Sedky, M

Digital DNA lifecycle security and privacy: An overview

<https://researchonline.ljmu.ac.uk/id/eprint/17149/>

Article

Citation (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

**Alsaffar, MM, Hasan, M, McStay, GP ORCID logoORCID:
<https://orcid.org/0000-0003-1363-8719> and Sedky, M (2022) Digital DNA lifecycle security and privacy: An overview. Briefings in Bioinformatics, 23 (2). ISSN 1467-5463**

LJMU has developed **LJMU Research Online** for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact researchonline@ljmu.ac.uk

1 Digital DNA Lifecycle Security and Privacy: An Overview

2 Alsaffar¹, M., Hasan¹, M., McStay², G.P., Sedky¹, M.

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

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

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

8

9 Key points

- 10 • The digital DNA life cycle describes all the processes and usages once the DNA has been
11 sequenced.
- 12 • One's privacy is threatened if their anonymised DNA is leaked; the threat level can be as high
13 as creating somebody's face image.
- 14 • Attempts to secure genomic data can fail or may not always scale to cover actual life data.
- 15 • A new approach powered by a Machine Learning (ML) solution is required to protect
16 genomic data.
17

18 Keywords:

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

22 Abstract

23 DNA sequencing technologies have advanced significantly in the last few years leading to
24 advancements in biomedical research which has improved personalised medicine and the discovery
25 of new treatments for diseases. Sequencing technology advancement has also reduced the cost of
26 DNA sequencing, which has led to the rise of Direct-To-Consumer (DTC) sequencing e.g.
27 23andme.com, ancestry.co.uk etc. In the meantime, concerns have emerged over privacy and
28 security in collecting, handling, analysing, and sharing DNA and genomic data.
29 DNA data is unique and can be used to identify individuals. Moreover, this data provides information
30 on people's current disease status and disposition e.g. mental health or susceptibility for developing
31 cancer. DNA privacy violation does not only affect the owner but also affects their close
32 consanguinity due to its hereditary nature.

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

38 1. Introduction

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

45 Genomic security is vital; if an adversary manages to gather one's genomic information, the
46 adversary would then be able to predict phenotypes such as facial structures. The ability to predict
47 physical traits and demographic information based on whole-genome sequences using Machine
48 Learning (ML) has advanced over the years [3]. Physical traits prediction is a significant threat to
49 privacy, and it also has important legal and ethical implications. The ability to predict physical traits
50 will also affect the suitability of current informed consent, the practicality and value of de-
51 identification of the supporting genomic information e.g. genomic owner's name and address [4].

52 Predicting facial structures based on whole-genome sequences has advanced even further. Research
53 by Qiao et al. [5] demonstrated that facial characteristics such as cheeks, mouth shape and other
54 facial features are related to as few as six genes and can be predicted from genomic data. Richmond
55 et al. [6] give a brief overview of the various facial genetics variants that influence facial phenotypes.

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

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

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

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

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

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

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

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

98 2. Digital DNA lifecycle

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

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

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

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

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

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

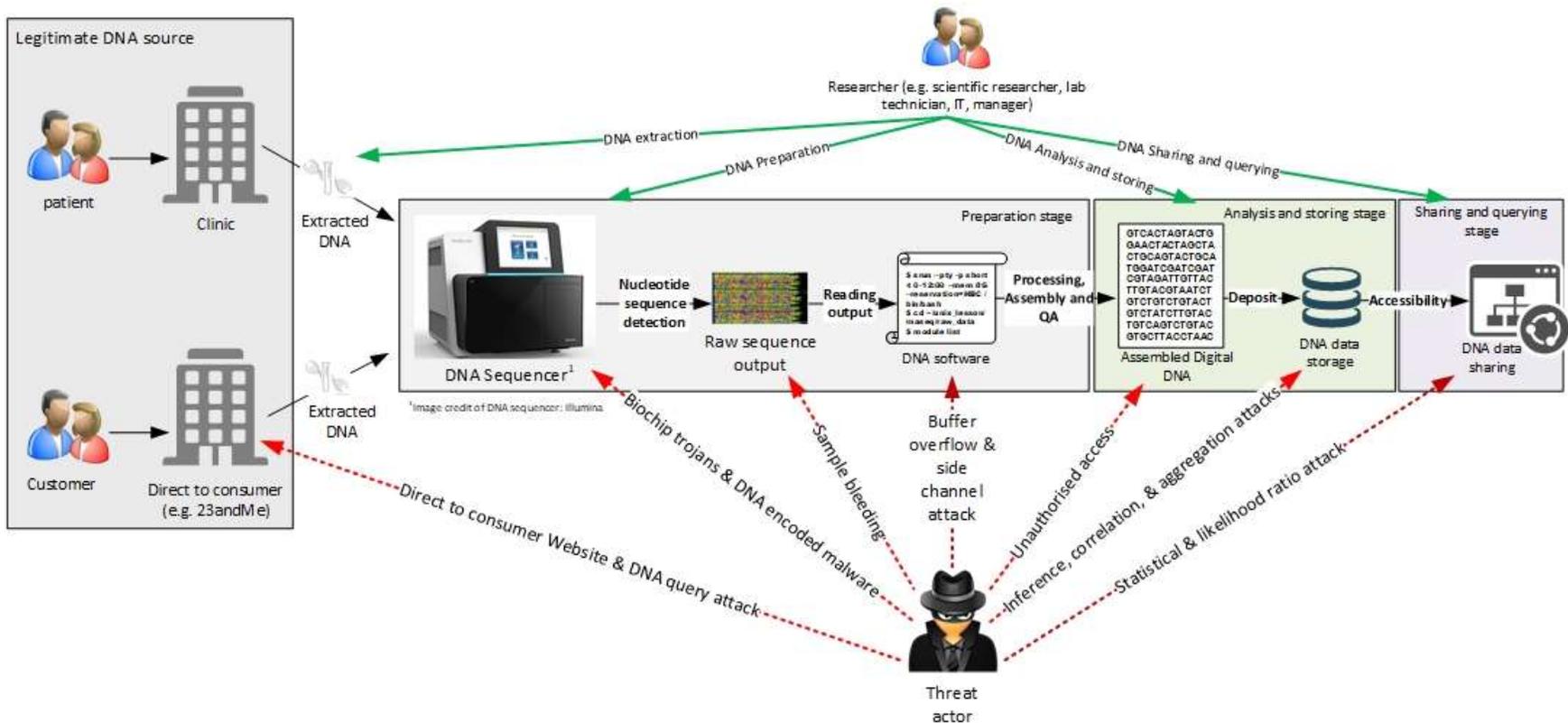


Figure 1. Digital DNA life cycle

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.

112 3. Preparation stage vulnerabilities

113 3.1. Encoding malware in a strand of DNA

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

117 3.1.1. Problem domain

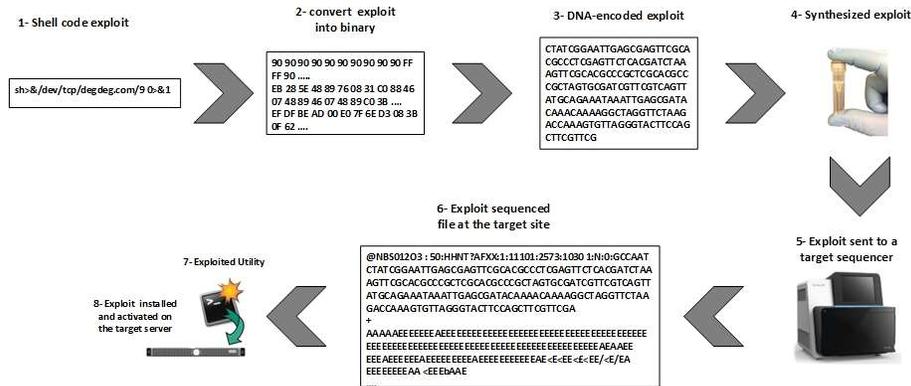
118 Ney et al. [30] demonstrated an adversary's ability to encode a malicious computer code into a
119 synthesised DNA sample. The authors were able to exploit a feature within the Linux operating
120 system which allowed them to receive a copy of all the network traffic generated in the DNA
121 alignment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

123 3.1.2. Available solutions

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

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

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



140

141 *Figure 2. DNA encoded malware adapted from [30]*

142 3.2. DNA sequencing hardware

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

144 3.2.1. Problem domain

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

153 Another vulnerability to DNA privacy stems from DNA sequencers hardware sequencing technology,
 154 as DNA sequencers commonly sequence thousands of samples from different sources
 155 simultaneously; this technique is known as multiplexing. Multiplexing relies on assigning unique 6 to
 156 8 digit identifiers to each sample; these identities can then be used to identify the sample during the
 157 demultiplexing process. The demultiplexing process (which is the process of separating the samples

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

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

163 3.2.2. Available solutions

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

168 3.3. DNA sequencing software

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

173 3.3.1. Problem domain

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

180 Ney et al. [30] assessed a sample software covering every stage of the DNA downstream pipeline.

181 The sample was grouped into specific categories and found to use many known insecure
182 functions/commands such as “strcpy” as shown in Table 1.

183 3.3.2. Available solutions
 184 Ney et al. [30] suggested that software security can be improved by following software security best
 185 practices including regular patching and updates.

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

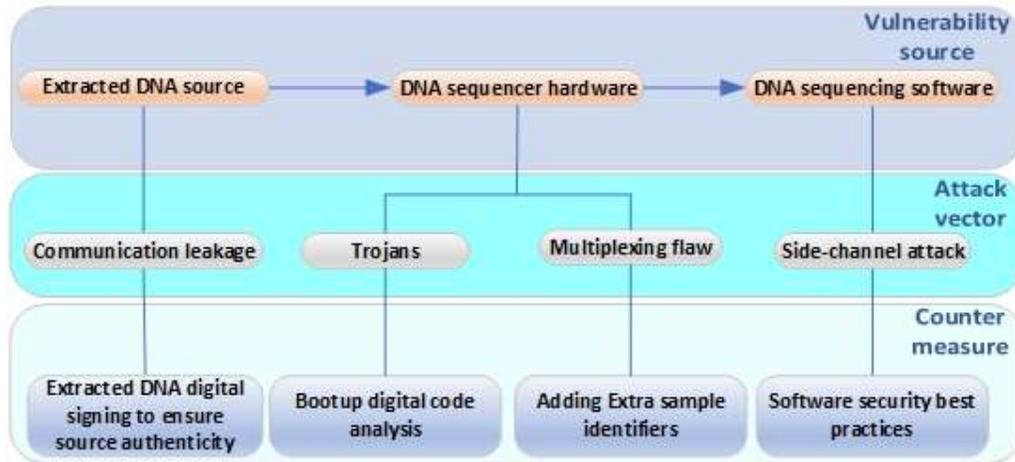
Category	Program	Version	Lines of Code	Normalised Count (Total Count)					
				strcat	strcpy	sprintf	vsprintf	gets	static buffers
NGS Analysis									
Preprocessing	fastx-toolkit	0.0.14	3189	0.314 (1)	0.314 (1)	0 (0)	0 (0)	0 (0)	14.425 (46)
	fqzcomp	4.6	2066	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	23.223 (48)
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)

188 3.4. Summary

189 DNA Sequencing hardware and software is vulnerable to misuse and errors (unintentional or
 190 otherwise). Figure 3 shows a summary of preparation stage vulnerabilities which consists of three
 191 levels; the top layer is the vulnerability source, the middle layer describes the attack vector and the
 192 bottom layer demonstrates the methods used to mitigate or reduce the risk of the attack vector.
 193 Extracted DNA sources can be tampered with to disrupt the sequencing cycle or create malware e.g.
 194 a worm that can infect the downstream computers and allow the attacker to receive a copy of their
 195 network communication. To reduce the risk, it is important to ensure that the DNA source is trusted
 196 and tracked which can be achieved by digitally signing DNA molecules.
 197 Two vulnerabilities reside in the DNA sequencer hardware i.e. trojans which can infect sequencer
 198 hardware, and the sequencers multiplexing flaw (sample bleed). Both flaws allow the attacker to
 199 infer or influence the sequenced samples. To reduce the risk of trojans, sequencers boot sequence

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

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



205

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

207

208 4. Analysis and storing stage vulnerabilities

209 4.1. DNA sequence read

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

214 4.1.1. Current solutions

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

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

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

228 4.1.2. Critical analysis

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

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

238 4.2. DNA alignment

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

244 4.2.1. Current solutions

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

251 Another option that is becoming more accessible is using a hybrid cloud. The speed on the hybrid
252 cloud has improved by utilising a secure Seed-and-extend read mapping algorithm. The algorithm
253 splits the computation such that the public cloud finds the exact seed matches using encrypted
254 seeds, and the private cloud extends the seed matches using unencrypted data [44]. The second
255 approach suggested by Popic et al. [45] is to preserve the read mapping's privacy for a hybrid cloud
256 using BALAUR. BALAUR preserves read mapper for hybrid cloud based on locality-sensitive hashing
257 and k-mer voting. It divides the computation between the trusted private client and the untrusted
258 public cloud. It operates in two phases; the first phase identifies a few candidates' positions in the
259 DNA strands where they can be aligned. These candidates are then assessed securely in the public
260 cloud against an already hashed and indexed dictionary that was pre-prepared using the private
261 client. This method is significantly faster than modern long read mappers, as the technique offloads
262 50–70% portion of the alignment to the cloud. However, Zhao et al. [46] created a new algorithm for
263 aligning short reads where encrypted data is aligned in the public cloud while encryption and
264 decryption occur in the private cloud. This algorithm produced results matching non-secure read
265 mapping.

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

271 [49] introduced a novel method called MASK AI alongside Intel SGX. The utility provides a two-tier
272 hybrid system; the first tier aligned masked reads in the public cloud while the second tier refines
273 the first tier results.

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

278 4.2.2. Critical analysis

279 BALAUR uses a lot of memory and requires substantial network bandwidth [49]. While Zhao et al.
280 [46] created an algorithm that works on short reads, most modern sequencers produce long reads
281 which might render this approach to be less beneficial for real-world usage. The use of Intel SGX is
282 limited by the size of the enclave which can vary depending on the number of processors and the
283 memory size [51]. This approach could work for a small data set; however, adding processing power
284 and memory can be proven to be costly if there is a need for a larger enclave in the cloud.

285 4.3. DNA data storage

286 Storing genomic data is the most common step after DNA alignment. However, there are no
287 standard rules to imply the retention and return policies and where to store the data which means
288 that research labs are expected to have their own standards. Most research labs store the genomic
289 data in the patients' medical records. Doing this may result in unintentional or malicious access by a
290 third party [52]. Vinatzer et al. [53] point out a lack of a mechanism to enforce adequate user
291 authentication. Most databases do not implement strong password requirements by default, and
292 access control is usually implemented when data is uploaded but rarely relevant when downloading
293 digital DNA data. Elgabry et al. [14] highlighted that an adversary could gain access to genomic
294 information by exploiting vulnerabilities within the database used to host the data. For example,
295 they can exploit database authentication weakness in MongoDB (the database used by Genomics
296 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.

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

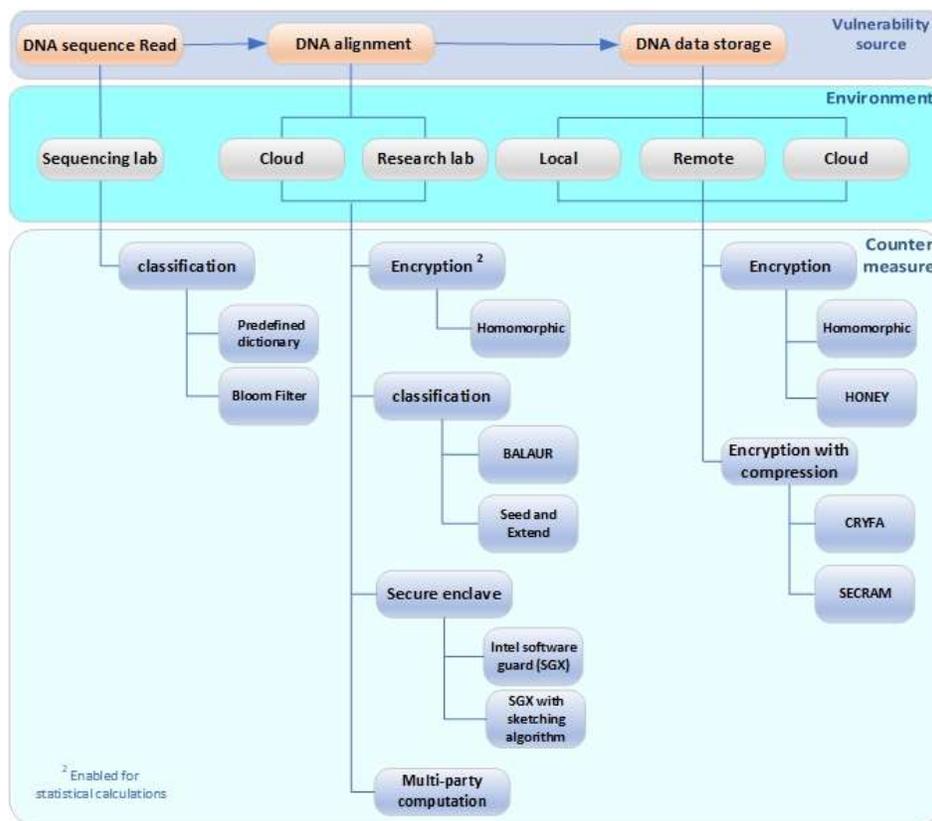
331 4.3.2. Critical analysis

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

339 4.4. Summary

340 The DNA analysis and storage stage which includes sequencing pipeline and post-sequencing storage
341 is at risk of unauthorised access and the disclosure of private information if the data is not
342 adequately protected. Therefore, researchers have utilised various methods to prevent
343 unauthorised data viewing. Figure 4 shows a summary of analysis and storing stage vulnerabilities. It
344 consists of three levels; the top layer is the vulnerability source, the middle layer is the environment
345 where the vulnerabilities can reside and the bottom layer is what can be done to mitigate or reduce
346 the risk of these vulnerabilities.
347 Sensitive DNA sequence reads can be viewed if not sufficiently protected. Sequence privacy can be
348 accomplished by using Classifications methods that classify reads into privacy-sensitive or non-

349 sensitive sections. Another approach is distributing the sequencing operation to multiple
 350 organisations, where each organisation will sequence a segment of the initial DNA.
 351 Four methods can be used to protect DNA privacy during alignment i.e. encryption, classification,
 352 secure enclave and multiparty computation. Memory randomisation or cache equalisation can hide
 353 access patterns to the reference DNA while aligning using a hybrid cloud.
 354 Encryption or encryption with compression and distributed storage can be used to preserve the
 355 privacy of the DNA data while stored (also known as data at rest) in a local, remote or cloud
 356 environment.



357
 358 *Figure 4. Summary of vulnerabilities associated with analysis and storing stage and their countermeasures*

359

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

361 5.1. Querying genomic data

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

364 5.1.1. Problem domain

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

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

376 5.1.2. Current solutions

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

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

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

396 5.2. Sharing genomic data

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

402 5.2.1. Data sharing limiting factors

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

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

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

419 5.2.2. Sharing standards

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

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

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

430 Another approach for genomic data sharing which can be used as a centralised approach is Genome-
431 Wide Association Study (GWAS) [75] or the decentralised approach which is the Federated GWAS
432 [76]. GWAS is set up to provide a repository with a large population to produce reliable statistical
433 results by using personal identifiable genetic markers. However, privacy concerns are making people
434 reluctant to contribute [77]. For researchers, genomic data provide an immense benefit if combined
435 with the patients' Electronic Health Records (EHRs). Hence, Harvard Medical School and the
436 Massachusetts Institute of Technology (MIT) developed Informatics for Integrating Biology and the
437 Bedside (i2b2) framework (now maintained by tranSMART Foundation). This solution can be
438 implemented on a single site [78], [79] or can combine data from multiple sites [80].

439 Each of these approaches has its limitations. Storing data in a centralised location will act as a single
440 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

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

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

480

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

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

490

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

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

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

513 Some researchers direct their efforts to secure specific fields in genomic studies e.g Gürsoy et al.
514 [97] introduced a new method to reduce private information leakage from functional genomics. The
515 researchers presented techniques to minimise common privacy risks that were quantified by
516 adopting statistical techniques. Jagadeesh et al. [98] provided a secure multiparty computation for
517 genomic diagnoses without revealing patient genomes based on two approaches. The first approach
518 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/

539 5.1. Direct-to-Consumer genetic testing

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

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

554 *Table 3: Popular Direct-to-Consumers (DTC) companies, the approximate customer numbers and the primary service*
555 *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	

556

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

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

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

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

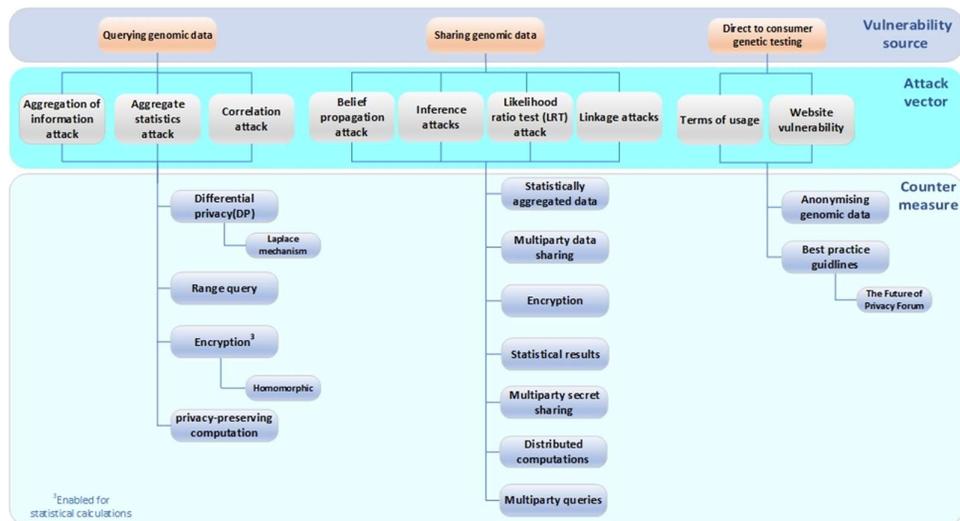
578 5.2. Summary

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

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

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

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



600

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

603 6. Conclusion

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

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

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

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

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

632 7. References

- 633 [1] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux, "De-anonymizing Genomic
634 Databases Using Phenotypic Traits," in *Proceedings on Privacy Enhancing Technologies*, 2015,
635 vol. 2.
- 636 [2] M. Backes, P. Berrang, M. Humbert, X. Shen, and V. Wolf, "Simulating the Large-Scale Erosion
637 of Genomic Privacy over Time," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 5, pp.
638 1405–1412, 2018.
- 639 [3] D. Sero *et al.*, "Facial recognition from DNA using face-to-DNA classifiers," *Nat. Commun.*, vol.
640 10, no. 1, pp. 1–12, Dec. 2019.
- 641 [4] C. Lippert *et al.*, "Identification of individuals by trait prediction using whole-genome
642 sequencing data.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 38, pp. 10166–10171, 2017.
- 643 [5] L. Qiao *et al.*, "Genome-wide variants of Eurasian facial shape differentiation and a
644 prospective model of DNA based face prediction," *J. Genet. Genomics*, vol. 45, no. 8, pp. 419–
645 432, 2018.
- 646 [6] S. Richmond, L. J. Howe, S. Lewis, E. Stergiakouli, and A. Zhurov, "Facial Genetics: A Brief
647 Overview," *Front. Genet.*, vol. 9, p. 462, 2018.
- 648 [7] M. Elgabry, D. Nesbeth, and S. D. Johnson, "A Systematic Review of the Criminogenic
649 Potential of Synthetic Biology and Routes to Future Crime Prevention," *Front. Bioeng.*
650 *Biotechnol.*, vol. 8, p. 1119, 2020.
- 651 [8] Y. Joly, I. N. Feze, L. Song, and B. M. Knoppers, "Comparative Approaches to Genetic

- 652 Discrimination: Chasing Shadows?," *Trends Genet.*, vol. 33, no. 5, pp. 299–302, 2017.
- 653 [9] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Quantifying Interdependent Risks in
654 Genomic Privacy," *ACM Trans. Priv. Secur.*, vol. 20, no. 1, pp. 1–31, Feb. 2017.
- 655 [10] E. Zeggini, A. L. Gloyn, A. C. Barton, and L. V Wain, "Translational genomics and precision
656 medicine: Moving from the lab to the clinic," *Science (80-.)*, vol. 365, no. 6460, pp. 1409 LP –
657 1413, Sep. 2019.
- 658 [11] Y. Y. Liu and S. A. Harbison, "A review of bioinformatic methods for forensic DNA analyses,"
659 *Forensic Science International: Genetics*, vol. 33. Elsevier Ireland Ltd, pp. 117–128, 01-Mar-
660 2018.
- 661 [12] N. Moray, K. E. Pink, P. Borry, and M. H. D. Larmuseau, "Paternity testing under the cloak of
662 recreational genetics," *Eur. J. Hum. Genet.*, vol. 25, no. 6, pp. 768–770, Mar. 2017.
- 663 [13] L. Vossaert, I. Chakchouk, R. Zemet, and I. B. Van den Veyver, "Overview and recent
664 developments in cell-based noninvasive prenatal testing," *Prenat. Diagn.*, vol. 41, no. 10, pp.
665 1202–1214, May 2021.
- 666 [14] M. Akgün, A. O. Bayrak, B. Ozer, and M. Ş. Sağıroğlu, "Privacy preserving processing of
667 genomic data: A survey," *J. Biomed. Inform.*, vol. 56, pp. 103–111, 2015.
- 668 [15] D. Lu *et al.*, "Methods of privacy-preserving genomic sequencing data alignments," *Brief.*
669 *Bioinform.*, May 2021.
- 670 [16] M. M. Al Aziz *et al.*, "Privacy-preserving techniques of genomic data-a survey," *Brief.*
671 *Bioinform.*, vol. 20, no. 3, pp. 887–895, 2017.
- 672 [17] A. Mittos, B. Malin, and E. De Cristofaro, "Systematizing Genome Privacy Research: A Privacy-
673 Enhancing Technologies Perspective," *Proc. Priv. Enhancing Technol.*, no. 1, pp. 87–107, 2019.
- 674 [18] X. Shi and X. Wu, "An overview of human genetic privacy," *Ann. N. Y. Acad. Sci.*, vol. 1387, no.
675 1, pp. 61–72, 2017.
- 676 [19] A. P. Schwab, H. S. Luu, J. Wang, and J. Y. Park, "Genomic Privacy," *Clin. Chem.*, vol. 64, no.
677 12, pp. 1696–1703, 2018.
- 678 [20] A. B. Carter, "Considerations for Genomic Data Privacy and Security when Working in the
679 Cloud," *J. Mol. Diagnostics*, vol. 21, no. 4, pp. 542–552, 2019.
- 680 [21] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature*
681 *Reviews Genetics*, vol. 15, no. 6. Nature Publishing Group, pp. 409–421, 2014.
- 682 [22] A. Mohammed Yakubu and Y.-P. P. P. Chen, "Ensuring privacy and security of genomic data
683 and functionalities," *Brief. Bioinform.*, vol. 21, no. 2, pp. 511–526, Mar. 2020.
- 684 [23] B. Abinaya and S. Santhi, "A survey on genomic data by privacy-preserving techniques
685 perspective," *Comput. Biol. Chem.*, vol. 93, p. 107538, Aug. 2021.
- 686 [24] M. Naveed *et al.*, "Privacy in the Genomic Era," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–44,
687 2015.
- 688 [25] A. Halimi *et al.*, "Privacy-Preserving and Efficient Verification of the Outcome in Genome-
689 Wide Association Studies," *ArXiv*, vol. abs/2101.0, 2021.
- 690 [26] L. Hartwell, M. Goldberg, J. Fischer, and L. Hood, *Genetics: from genes to genomes*. McGraw-
691 Hill Education, 2018.
- 692 [27] K. L. Howe *et al.*, "Ensembl 2021," *Nucleic Acids Res.*, vol. 49, no. 2, 2021.

- 693 [28] V. Jalili, M. Matteucci, M. Masseroli, and S. Ceri, "Indexing Next-Generation Sequencing
694 data," *Inf. Sci. (Ny)*, vol. 384, pp. 90–109, Apr. 2017.
- 695 [29] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, "Demonstration of end-to-end
696 Automation of DNA Data storage," *Sci. Rep.*, vol. 9, no. 1, pp. 1–5, 2019.
- 697 [30] P. Ney, K. Koscher, L. Organick, L. Ceze, and T. Kohno, "Computer security, privacy, and DNA
698 sequencing: compromising computers with synthesized DNA, privacy leaks, and more," in
699 *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 765–779.
- 700 [31] J. E. Gallegos, D. M. Kar, I. Ray, I. Ray, and J. Peccoud, "Securing the Exchange of Synthetic
701 Genetic Constructs Using Digital Signatures," *ACS Synth. Biol.*, vol. 9, no. 10, pp. 2656–2664,
702 Oct. 2020.
- 703 [32] S. Allen Morse, G. D. Koblenz, J. Diggans, and E. Leproust, "POLICY AND PRACTICE REVIEWS
704 Next Steps for Access to Safe, Secure DNA Synthesis," *Front. Bioeng. Biotechnol.*, vol. 1, p. 86,
705 2019.
- 706 [33] S. S. Ali, M. Ibrahim, J. Rajendran, O. Sinanoglu, and K. Chakrabarty, "Supply-Chain Security of
707 Digital Microfluidic Biochips," *Computer (Long. Beach. Calif.)*, vol. 49, no. 8, pp. 36–43, Aug.
708 2016.
- 709 [34] I. Fayans, Y. Motro, L. Rokach, Y. Oren, and J. Moran-Gilad, "Cyber security threats in the
710 microbial genomics era: implications for public health," *Eurosurveillance*, vol. 25, no. 6, p.
711 1900574, Feb. 2020.
- 712 [35] Q. Li *et al.*, "Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS
713 platform," *BMC Genomics*, vol. 20, no. 1, p. 215, 2019.
- 714 [36] A. Mitra, M. Skrzypczak, K. Ginalski, and M. Rowicka, "Strategies for Achieving High
715 Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina
716 Platform," *PLoS One*, vol. 10, no. 4, p. 120520, Apr. 2015.
- 717 [37] S. Arshad, J. Arshad, M. M. Khan, and S. Parkinson, "Analysis of security and privacy
718 challenges for DNA-genomics applications and databases," *J. Biomed. Inform.*, vol. 119, p.
719 103815, Jul. 2021.
- 720 [38] J. Decouchant, M. Fernandes, M. Völp, F. M. Couto, and P. Esteves-Verissimo, "Accurate
721 filtering of privacy-sensitive information in raw genomic data," *J. Biomed. Inform.*, vol. 82, pp.
722 1–12, 2018.
- 723 [39] V. V. Cogo, A. Bessani, F. M. Couto, and P. Verissimo, "A High-Throughput Method to Detect
724 Privacy-Sensitive Human Genomic Data," in *Proceedings of the 14th ACM Workshop on*
725 *Privacy in the Electronic Society - WPES '15*, 2015, pp. 101–110.
- 726 [40] M. Fernandes, J. Decouchant, M. Volp, F. M. Couto, and P. Esteves-Verissimo, "DNA-SeAI:
727 Sensitivity Levels to Optimize the Performance of Privacy-Preserving DNA Alignment," *IEEE J.*
728 *Biomed. Heal. Informatics*, vol. 24, no. 3, pp. 907–915, Mar. 2020.
- 729 [41] A. Gholami, M. A. Maddah-Ali, and S. Abolfazl Motahari, "Private Shotgun DNA Sequencing,"
730 2019, vol. 2019-July, pp. 171–175.
- 731 [42] M. Z. Hasan, M. S. R. Mahdi, M. N. Sadat, and N. Mohammed, "Secure count query on
732 encrypted genomic data," *J. Biomed. Inform.*, vol. 81, pp. 41–52, 2018.
- 733 [43] S. Roy *et al.*, "Standards and Guidelines for Validating Next-Generation Sequencing
734 Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology
735 and the College of American Pathologists," *J. Mol. Diagnostics*, vol. 20, no. 1, pp. 4–27, 2018.
- 736 [44] V. Popic and S. Batzoglou, "Privacy-Preserving Read Mapping Using Locality Sensitive Hashing

- 737 and Secure Kmer Voting,” *bioRxiv*, p. 46920, Jan. 2016.
- 738 [45] V. Popic and S. Batzoglou, “A hybrid cloud read aligner based on MinHash and kmer voting
739 that preserves privacy,” *Nat. Commun.*, vol. 8, no. 1, p. 15311, May 2017.
- 740 [46] Y. Zhao, X. Wang, and H. Tang, “A Secure Alignment Algorithm for Mapping Short Reads to
741 Human Genome,” *J. Comput. Biol.*, vol. 25, no. 6, pp. 529–540, May 2018.
- 742 [47] F. Chen *et al.*, “PRESAGE: PRivity-preserving gEnetic testing via SoftwAre Guard Extension,”
743 *BMC Med. Genomics*, vol. 10, no. S2, p. 48, 2017.
- 744 [48] C. Kockan *et al.*, “Sketching algorithms for genomic data analysis and querying in a secure
745 enclave,” *Nat. Methods*, vol. 17, no. 3, pp. 295–301, 2020.
- 746 [49] C. Lambert, M. Fernandes, J. Decouchant, and P. Esteves-Verissimo, “MaskAI: Privacy
747 Preserving Masked Reads Alignment using Intel SGX,” in *2018 IEEE 37th Symposium on
748 Reliable Distributed Systems (SRDS)*, 2018, pp. 113–122.
- 749 [50] M. Völz, J. Decouchant, C. Lambert, M. Fernandes, and P. Esteves-Verissimo, “Enclave-Based
750 Privacy-Preserving Alignment of Raw Genomic Information,” in *Proceedings of the 2nd
751 Workshop on System Software for Trusted Execution - SysTEX’17*, 2017, pp. 1–6.
- 752 [51] “Intel® Software Guard Extensions (Intel® SGX).” [Online]. Available:
753 [https://www.intel.co.uk/content/www/uk/en/architecture-and-technology/software-guard-
754 extensions.html](https://www.intel.co.uk/content/www/uk/en/architecture-and-technology/software-guard-extensions.html). [Accessed: 19-Oct-2021].
- 755 [52] M. Shabani, D. Vears, and P. Borry, “Raw Genomic Data: Storage, Access, and Sharing,”
756 *Trends Genet.*, vol. 34, no. 1, pp. 8–10, Jan. 2018.
- 757 [53] B. A. Vinatzer, L. S. Heath, H. M. J. Almohri, M. J. Stulberg, C. Lowe, and S. Li,
758 “Cyberbiosecurity Challenges of Pathogen Genome Databases,” *Front. Bioeng. Biotechnol.*,
759 vol. 7, p. 106, 2019.
- 760 [54] Genomics England, “Genomics England,” 2021. [Online]. Available:
761 <https://www.genomicsengland.co.uk/>. [Accessed: 06-Feb-2021].
- 762 [55] Z. Huang *et al.*, “A privacy-preserving solution for compressed storage and selective retrieval
763 of genomic data,” *Genome Res.*, vol. 26, no. 12, pp. 1687–1696, 2016.
- 764 [56] D. Hwang, S. Choi, J. Shin, G. Song, and Y. Choi, “Privacy-Preserving Compressed Reference-
765 Oriented Alignment Map Using Decentralized Storage,” *IEEE Access*, vol. 6, pp. 45990–46001,
766 2018.
- 767 [57] J.-P. Aumasson, “The impact of quantum computing on cryptography,” *Comput. Fraud Secur.*,
768 vol. 2017, no. 6, pp. 8–11, 2017.
- 769 [58] M. Hosseini, D. Pratas, and A. J. Pinho, “Cryfa: a secure encryption tool for genomic data,”
770 *Bioinformatics*, vol. 35, no. 1, pp. 146–148, 2019.
- 771 [59] Z. Huang, E. Ayday, J. Fellay, J.-P. Hubaux, and A. Juels, “GenoGuard: Protecting Genomic Data
772 against Brute-Force Attacks,” in *2015 IEEE Symposium on Security and Privacy*, 2015, pp. 447–
773 462.
- 774 [60] J. S. Sousa *et al.*, “Efficient and secure outsourcing of genomic data storage,” *BMC Med.
775 Genomics*, vol. 10, no. 2, pp. 15–28, 2017.
- 776 [61] L. Chen, M. M. Aziz, N. Mohammed, and X. Jiang, “Secure large-scale genome data storage
777 and query,” *Comput. Methods Programs Biomed.*, vol. 165, pp. 129–137, Oct. 2018.
- 778 [62] N. Almadhoun, E. Ayday, and Ö. Ulusoy, “Inference attacks against differentially private query

- 779 results from genomic datasets including dependent tuples," *Bioinformatics*, vol. 36, no.
780 Supplement_1, pp. i136–i145, Jul. 2020.
- 781 [63] S. S. Samani *et al.*, "Quantifying Genomic Privacy via Inference Attack with High-Order SNV
782 Correlations," in *2015 IEEE Security and Privacy Workshops*, 2015, pp. 32–40.
- 783 [64] J. Henriksen-Bulmer and S. Jeary, "Re-identification attacks—A systematic literature review,"
784 *Int. J. Inf. Manage.*, vol. 36, no. 6, Part B, pp. 1184–1192, 2016.
- 785 [65] B. Wang, W. Song, W. Lou, and Y. T. Hou, "Privacy-preserving pattern matching over
786 encrypted genetic data in cloud computing," in *Proceedings - IEEE INFOCOM*, 2017, pp. 1–9.
- 787 [66] X. Ding, E. Ozturk, and G. Tsudik, "Balancing security and privacy in genomic range queries,"
788 in *Proceedings of the ACM Conference on Computer and Communications Security*, 2019, pp.
789 106–110.
- 790 [67] W. Briguglio, P. Moghaddam, W. A. Yousef, I. Traoré, and M. Mamun, "Machine learning in
791 precision medicine to preserve privacy via encryption," *Pattern Recognit. Lett.*, vol. 151, pp.
792 148–154, 2021.
- 793 [68] T. Haeusermann, B. Greshake, A. Blasimme, D. Irdam, M. Richards, and E. Vayena, "Open
794 sharing of genomic data: Who does it and why?," *PLoS One*, vol. 12, no. 5, p. e0177158, 2017.
- 795 [69] E. Ayday and M. Humbert, "Inference Attacks against Kin Genomic Privacy," *IEEE Secur. Priv.*,
796 vol. 15, no. 5, pp. 29–37, 2017.
- 797 [70] B. Berger and H. Cho, "Emerging technologies towards enhancing privacy in genomic data
798 sharing," *Genome Biol.*, vol. 20, no. 1, p. 128, 2019.
- 799 [71] Genomics England, "Genome Sequencing," 2021. [Online]. Available:
800 <https://www.genomicsengland.co.uk/understanding-genomics/genome-sequencing/>.
801 [Accessed: 25-Apr-2021].
- 802 [72] A. Page *et al.*, "A federated ecosystem for sharing genomic, clinical data," *Am. Assoc. Adv.*
803 *Sci.*, vol. 352, no. 6291, pp. 1278 LP – 1280, Jun. 2016.
- 804 [73] GA4GH, "BEACON," 2021. [Online]. Available: <https://beacon-project.io/>. [Accessed: 14-Aug-
805 2020].
- 806 [74] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, "Aftermath of bustamante
807 attack on genomic beacon service," *BMC Med. Genomics*, vol. 10, no. S2, p. 43, 2017.
- 808 [75] T. Beck, T. Shorter, and A. J. Brookes, "GWAS Central: a comprehensive resource for the
809 discovery and comparison of genotype and phenotype data from genome-wide association
810 studies," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D933–D940, Jan. 2020.
- 811 [76] X. Wu *et al.*, "A novel privacy-preserving federated genome-wide association study
812 framework and its application in identifying potential risk variants in ankylosing spondylitis,"
813 *Brief. Bioinform.*, vol. 22, no. 3, May 2021.
- 814 [77] C. Bonte, E. Makri, A. Ardehirdavani, J. Simm, Y. Moreau, and F. Vercauteren, "Towards
815 practical privacy-preserving genome-wide association study," *BMC Bioinformatics*, vol. 19, no.
816 1, p. 537, 2018.
- 817 [78] J. L. Raisaro *et al.*, "Protecting Privacy and Security of Genomic Data in i2b2 With
818 Homomorphic Encryption and Differential Privacy," *IEEE/ACM Trans. Comput. Biol.*
819 *Bioinforma.*, pp. 1–1, 2018.
- 820 [79] TranSMART, "i2b2," 2021. [Online]. Available: <https://www.i2b2.org/>. [Accessed: 17-Jul-
821 2021].

- 822 [80] J. G. Klann, A. Abend, V. A. Raghavan, K. D. Mandl, and S. N. Murphy, "Data interchange using
823 i2b2," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 5, pp. 909–915, 2016.
- 824 [81] J. L. Raisaro *et al.*, "MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed
825 Clinical and Genomic Data," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 16, no. 4, pp.
826 1328–1341, 2018.
- 827 [82] R. Cai *et al.*, "Deterministic identification of specific individuals from GWAS results,"
828 *Bioinformatics*, vol. 31, no. 11, pp. 1701–1707, Jun. 2015.
- 829 [83] Z. He, J. Yu, J. Li, Q. Han, G. Luo, and Y. Li, "Inference Attacks and Controls on Genotypes and
830 Phenotypes for Individual Genomic Data," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol.
831 17, no. 3, pp. 1–1, May 2018.
- 832 [84] Y. Wang, J. Wen, X. Wu, and X. Shi, "Infringement of Individual Privacy via Mining
833 Differentially Private GWAS Statistics," in *International Conference on Big Data Computing
834 and Communications*, 2016, pp. 355–366.
- 835 [85] L. Zhang, Q. Pan, Y. Wang, X. Wu, and X. Shi, "Bayesian network construction and genotype-
836 phenotype inference using GWAS statistics," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol.
837 16, no. 2, pp. 475–489, Mar. 2019.
- 838 [86] Y. Zhang, M. Blanton, and G. Almashaqbeh, "Secure distributed genome analysis for GWAS
839 and sequence comparison computation," *BMC Med. Inform. Decis. Mak.*, vol. 15 Suppl 5, no.
840 S5, pp. S4–S4, 2015.
- 841 [87] Z. Wan, Y. Vorobeychik, E. W. Clayton, M. Kantarcioglu, and B. Malin, "Game theory for
842 privacy-preserving sharing of genomic data," in *Responsible Genomic Data Sharing*, X. Jiang
843 and H. B. T.-R. G. D. S. Tang, Eds. Academic Press, 2020, pp. 135–160.
- 844 [88] T. Pascoal, J. Decouchant, A. Boutet, and P. Esteves-Verissimo, "DyPS: Dynamic, Private and
845 Secure GWAS," in *Proceedings on Privacy Enhancing Technologies*, 2021, vol. 2021, no. 2, pp.
846 214–234.
- 847 [89] M. Wang *et al.*, "Mechanisms to protect the privacy of families when using the transmission
848 disequilibrium test in genome-wide association studies," *Bioinformatics*, vol. 33, no. 23, pp.
849 3716–3725, Dec. 2017.
- 850 [90] N. Von Thenen, E. Ayday, and A. E. Cicek, "Re-identification of individuals in genomic data-
851 sharing beacons via allele inference," *Bioinformatics*, vol. 35, no. 3, pp. 365–371, 2019.
- 852 [91] J. L. Raisaro *et al.*, "Addressing Beacon re-identification attacks: Quantification and mitigation
853 of privacy risks," *J. Am. Med. Informatics Assoc.*, vol. 24, no. 4, 2017.
- 854 [92] D. Demmler, K. Hamacher, T. Schneider, and S. Stammer, "Privacy-preserving whole-genome
855 variant queries," in *6th International Conference on Cryptology and Network Security*, 2018,
856 vol. 11261 LNCS, pp. 71–92.
- 857 [93] F. Chen *et al.*, "PRINCESS: Privacy-protecting Rare disease International Network
858 Collaboration via Encryption through Software guard extensionS," *Bioinformatics*, vol. 33, no.
859 6, pp. 871–878, Mar. 2017.
- 860 [94] T. Schneider and O. Tkachenko, "Towards Efficient Privacy-Preserving Similar Sequence
861 Queries on Outsourced Genomic Databases," in *Proceedings of the 2018 Workshop on Privacy
862 in the Electronic Society*, 2018, pp. 71–75.
- 863 [95] H. I. Ozercan, A. M. Ileri, E. Ayday, and C. Alkan, "Realizing the potential of blockchain
864 technologies in genomics," *Genome Res.*, vol. 28, no. 9, pp. 1255–1263, 2018.
- 865 [96] D. Grishin, K. Obbad, and G. M. Church, "Data privacy in the age of personal genomics," *Nat.*

- 866 *Biotechnol.*, vol. 37, no. 10, pp. 1115–1117, 2019.
- 867 [97] G. Gürsoy *et al.*, “Data Sanitization to Reduce Private Information Leakage from Functional
868 Genomics,” *Cell*, vol. 183, no. 4, pp. 905–917.e16, 2020.
- 869 [98] K. A. Jagadeesh *et al.*, “Deriving genomic diagnoses without revealing patient genomes,”
870 *Science (80-.)*, vol. 357, no. 6352, pp. 692–695, Aug. 2017.
- 871 [99] M. Akgün, A. B. Ünal, B. Ergüner, N. Pfeifer, and O. Kohlbacher, “Identifying disease-causing
872 mutations with privacy protection,” *Bioinformatics*, vol. 36, no. 21, pp. 5205–5213, Nov.
873 2020.
- 874 [100] J. Kelleher *et al.*, “Htsget: A protocol for securely streaming genomic data,” *Bioinformatics*,
875 vol. 35, no. 1, pp. 119–121, 2019.
- 876 [101] L. Bonomi, Y. Huang, and L. Ohno-Machado, “Privacy challenges and research opportunities
877 for genomic data sharing,” *Nat. Genet.*, vol. 52, pp. 646–654, 2020.
- 878 [102] L. I. Laestadius, J. R. Rich, and P. L. Auer, “All your data (effectively) belong to us: Data
879 practices among direct-to-consumer genetic testing firms,” *Genet. Med.*, vol. 19, pp. 513–
880 520, 2017.
- 881 [103] L. Defrancesco and A. Klevecz, “Your DNA broker,” *Nat. Biotechnol.*, vol. 37, no. 10, pp. 842–
882 847, 2019.
- 883 [104] P. Ney, L. Ceze, and T. Kohno, “Genotype extraction and false relative attacks: security risks to
884 third-party genetic genealogy services beyond identity inference,” in *Network and Distributed
885 System Security Symposium (NDSS)*, 2020.
- 886 [105] The Future of PrivacyForum, “The Future of Privacy Forum.” [Online]. Available:
887 <https://fpf.org/>. [Accessed: 20-Apr-2021].
- 888 [106] R. M. Hendricks-Sturup and C. Y. Lu, “Direct-to-Consumer Genetic Testing Data Privacy: Key
889 Concerns and Recommendations Based on Consumer Perspectives,” *J. Pers. Med.*, vol. 9, no.
890 2, 2019.
- 891 [107] M. G. Hansson *et al.*, “The risk of re-identification versus the need to identify individuals in
892 rare disease research,” *Eur. J. Hum. Genet.*, vol. 24, no. 11, pp. 1553–1558, 2016.
- 893