Security, privacy, and trust management in DNA computing

Fernandes, Maria; Decouchant, Jérémie; Couto, Francisco M.

DOI
10.1016/bs.adcom.2022.08.009

Publication date
2023

Document Version
Proof

Published in
Perspective of DNA Computing in Computer Science

CHAPTER TWO

Security, privacy, and trust management in DNA computing

Maria Fernandes[a,*], **Jérémie Decouchant**[b], **and Francisco M. Couto**[c]

[a]SnT-University of Luxembourg, Esch-sur-Alzette, Luxembourg
[b]Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands
[c]LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

Contents

Abstract

DNA computing is an emerging field that aims at enabling more efficient data storage and processing. One principle of DNA computing is to encode some information

*Current affiliation: Big Data Institute, University of Oxford, Oxford, United Kingdom; Wellcome Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

Advances in Computers, Volume 129
ISSN 0065-2458
https://doi.org/10.1016/bs.adcom.2022.08.009

(e.g., image, video, programming scripts) into a digital DNA-like sequence and then synthesize the corresponding DNA molecule. Synthesizing this molecule using digital or real human genomic fragments theoretically opens the possibility for privacy attacks, which have been demonstrated on a large array of human genomic data. These privacy attacks aim at breaching the privacy of DNA samples, allowing an attacker to discover privacy-critical information from the partial or complete DNA information of an individual. In the context of DNA computing, novel privacy attacks will certainly emerge and could consist in discovering a part of a particular script or video that is privacy-critical. It is therefore important to consider whether privacy attacks and defense mechanisms can be used when manipulating genomic data. First, this chapter provides the background about genomic data, and its modern generation and processing. It then provides a survey on known genomic privacy attacks, and presents the privacy-enhancing technologies that have been designed to protect genomic data. Later, this chapter also introduces the current trust management methods one can rely on to further secure DNA storage and processing methods, before discussing how DNA computing currently relates to those attacks and privacy-preserving technologies. Finally, this chapter presents future research avenues.

Abbreviations

1000 GP	1000 Genomes Project
AES	Advanced Encryption Standard
CSP	Cloud Service Provider
dbGAP	database of Genotypes and Phenotypes
DNA	Desoxyribonucleic Acid
GA4GH	Global Alliance for Human Genetics
GWASs	Genome-Wide Association Studies
IBS	Identical-By-State
LD	Linkage Disequilibrium
NGS	Next Generation Sequencing
SGX	Software Guard Extensions
SMPC	Secure Multiparty Computations
SNPs	Single Nucleotide Polymorphisms
TEEs	Trusted Execution Environments
WGS	Whole Genome Sequencing

1. Introduction

Human Deoxyribonucleic Acid (DNA) is the genetic material contained in human cells that encode all the information necessary for an organism's functioning and reproduction. Therefore, understanding how DNA modulates all those processes became a hot topic due to its potential contributions on fields such as healthcare and forensics.

The DNA molecule is composed of four nucleotides, i.e., adenine (A), thymine (T), guanine (G), and cytosine (C), that bind together to form a double helix. The two strands that compose the double helix are antiparallel,

and, for each individual, contain genomic variations at certain positions in the genome. Genomic variations are alternative genomic subsequences where individuals might differ, the most common among people are Single Nucleotide Polymorphisms (SNPs). SNPs are genomic variation where there is a single nucleotide difference, either because a nucleotide has been replaced by another one, or because a nucleotide has been inserted or deleted. SNPs are also the most studied genomic variations because of their low complexity. Genomic variations among and between individuals are interconnected by complex statistical relations, such as Linkage Disequilibrium (LD) and kinship, which respectively describe dependencies between different regions in the genome, and between genomes from relatives. LD describes the nonrandom associations of a group of genomic variations, which comes from their simultaneous transmission during cell divisions. The resulting statistical associations can be exploited to infer a genomic variation when others are known. *Kinship* or kin relationships describe hereditary connections and marriage ties between individuals, including direct bonds (e.g., children, parents, grandparents) and collateral bonds (e.g., siblings, cousins, aunts, uncles). Kin relationships result in genomic similarities between members of the same family because of the biological inheritance process that dictates the transmission of genetic information from parents to their children.

The notions of genotypes and phenotypes are important to understand the genomic data field, and are defined and correlated as follows. An individual's genotype corresponds to its set of genomic variations. The phenotype of an individual is the set of its observable physical characteristics, such as its appearance (e.g., skin color/type, hair color/type, body silhouette), development (e.g., blood cells, hormones production), and behavior. The phenotype is the result of the expression of the information in the genotype in combination with the environment interactions (i.e., epigenetics). Since the environment has a great impact on the phenotype, even individuals with similar genotypes, such as twins, can present different phenotypes.

The inclusion of genomic data in multiple scientific areas was promoted by the advances of Next Generation Sequencing (NGS) technologies, which decreased the data generation cost and, consequently, increased the availability of genomic data. The first step to reveal the DNA information encoded by a sample is to sequence it using a NGS technology. NGS technologies are machines that perform a chain of chemical reactions on a biological sample (e.g., a blood sample) to translate it into its digital equivalent, in the form of sequences of nucleotide called reads. After sequencing, the information retrieved is treated in a processing pipeline to identify its special features, i.e., its genomic variations.

Let us briefly introduce the main steps of this processing pipeline: read alignment and variant calling. Read alignment is the first step required to determine the biological information contained on the reads produced by NGS technologies. In this step the reads are mapped to a reference genome to determine their original position in the genome. The reference genome is a synthetic genome sequence containing the most common genomic variations in the global population. The human reference genome was assembled based on the genome of several individuals from all around the world. Therefore, it represents a global synthetic sequence and not a single individual's genome. Variant calling is the step where the aligned reads are compared to the reference genome to identify the positions at which they differ. In this step, a quality score is used to distinguish from real genomic variations and sequencing errors.

DNA data have been increasingly used in healthcare data processing pipelines, such as personalized medicine and disease predisposition testing, research, direct-to-consumer services, and forensics [1]. The high throughput of sequencing machines, which encourages huge DNA data production, and the intensive computations required to process genomic data, often leads these processing pipelines to be outsourced to cloud environments that provide powerful computational resources at an affordable cost. Cloud-based environments for biomedical data have been described in Refs. [2, 3]. However, although public clouds provide powerful computational resources at an affordable cost, they are managed by a third party, i.e., a Cloud Service Provider (CSP). Using public cloud therefore raises new challenges in order to keep genomic data secure [4]. Furthermore, the increasing availability of genomic data and its large array of potential applications encouraged data sharing to accelerate the understanding of DNA functioning, and allow the largest number to benefit from the information it encodes. Due to the important expected applications of DNA data, and because of its greater availability, several genomic data repositories were created to support and speed up knowledge acquisition and creation. Some examples of those repositories include the 1000 Genomes Project (1000 GP), the database of Genotypes and Phenotypes (dbGaP), and the 100,000 Genomes projects. The 1000 GP is a publicly available repository of human genomes, launched in 2008 with the goal of creating a resource on human genetic variations. dbGaP is a repository containing genotype and phenotype data, which is an important collection for Genome-Wide Association Studies (GWASs), and genome-diseases correlation studies. The 100,000 Genomes Project is an England effort to provide a repository of cancer and rare diseases

genomic data to boost research in these fields. GWASs are a particular DNA study whose goal is to link observed genomic variations with particular diseases. GWAS consists of a massive scan over multiple individuals' genomes to search for particular patterns that help to predict occurrences of a disease. Once those patterns are identified, they can be used to study the contribution of genes to the disease, and improve its diagnostic and treatment. Along with the increase of genomic data processing, new requirements have emerged such as the needs for high performance and privacy. The high-performance demand pushed for the use of scalable and cost-efficient environments, such as public clouds, that provide powerful computational resources and large storage capacity. However, as DNA data is directly linked to its owner identity, privacy breaches can occur when data is not protected before it is sent to the cloud.

The emphasis put on genomic data security has been growing with the application of genomic data in developing fields. Enforcing data security in an information system requires providing both privacy and trust. Privacy challenges may arise when genomic data, which is sensitive information, is stored and processed in a cloud environment, or shared. The privacy risks that appear when outsourcing biomedical data to public clouds without adequate protections are discussed in Refs. [5, 6]. Genomic data carry sensitive information such as predisposition to genetic diseases, physical traits and familial relations. As have seen described, members of a family share genomic traits, and genome correlations. Humbert *et al.* [7] demonstrate that the genomic privacy of a target individual decreases when genomic information from its family members is shared. This work highlights the increasing of privacy risks, since human genomes sequencing is constantly growing. In addition, the nonrevocable nature of DNA makes any potential data leakage result in privacy loss that can never be attenuated. Therefore, such data should be kept secret to prevent any harm to the owner, such as genetic discrimination, which could result in denial of health insurance, education, and employment, or blackmail [8]. From a security perspective, biological correlations between human genomes should also be considered when designing DNA data processing algorithms, since they can be exploited by an adversary to infer further information based on a partial genomic sequence and known statistics and/or to relate family members. Therefore, it is important to protect family relationship information to ideally prevent privacy attacks. A deeper discussion on privacy attacks on genomic data and existing privacy-preserving techniques can be found in Ref. [9].

DNA computing is an emerging field that aims at using DNA to store information to perform computations through chemical reactions. For this purpose, synthesized DNA is produced, by first composing DNA sequences in silico and then producing the corresponding DNA molecules in laboratory. Since each molecule is synthesized to store digital information, i.e., video, photo, code, its sequence is obtained by using a binary correspondence for each nucleotide. For example, a basic encoding of nucleotides over 2 bits could be: A = [00], T = [01], C = [10], and G = [11]. DNA computing promises the development of massive parallel computing technologies, which would allow complex problems to be solved in a short amount of time, instead of requiring weeks using conventional computers. These promises rely on the fact that millions of DNA molecules can interact simultaneously. However, this also increases the complexity of the output that a DNA computer would provide. Human DNA and synthesized DNA (generated for DNA computing) share the same structure; however, since they encode different information, they can have different properties. Yet, an interesting practical application of DNA computing, in which the synthesized DNA has similar properties as the DNA found in the human body, is the synthesis of DNA molecules to detect cancerous or damaged cells with the goal of triggering the repairing response on them. This process allows the prevention of the rapid multiplication of such cells and therefore slows the effects of the resulting illnesses. This process has been described by Shapiro *et al.* [10]. In this context, the synthesized DNA generated, which uses as template the human DNA, presents the statistical correlations it possesses such as LD and kinship. Therefore, similar to the sequenced human DNA, the synthesized DNA is vulnerable to security attacks performed on human DNA. Such attacks aim at disclosing sensitive information about the owner based on his/her DNA sequence or on the DNA sequence of its relatives. This chapter focuses on the security, privacy, and trust management aspect of human DNA. The goal of this chapter is to show the privacy risks that synthesized DNA and human DNA share. We therefore describe security attacks that have been performed on the latter and provide an overview of the state-of-the-art techniques one can use to prevent such attacks. In addition, this chapter also provides guidelines to maintain trust when designing privacy-preserving solutions for human-like DNA. Overall, this chapter makes the following contributions.

1. First, this chapter describes the security and privacy challenges in the context of human DNA and DNA computing, and put the emphasis on existing privacy attacks on genomic data.
2. Second, it surveys the scientific community's efforts to develop privacy-preserving techniques for genomic data.

3. Third, it highlights how trust can be maintained while processing genomic data.
4. Finally, this chapter discusses the relations and impact of privacy-preserving techniques and genomic data privacy attacks on DNA computing.

The remainder of this chapter is organized as follows. Section 2 provides an overview on the existing security attacks on genomic data. Section 3 presents the community effort to develop privacy-preserving solutions to prevent the reported security attacks. Section 4 describes the important trust management aspects for the design of secure genomic data processing and storage solutions. Section 5 discusses the previous sections, and the evolution of the scientific community best practices. Section 6 concludes this chapter and provides some insights for future work.

2. Security attacks

Over the past decades, many privacy attacks on genomic data have been reported. These attacks explored genomic data features obtained from a single individual, from a family, or from a target group. These attacks alerted the research community of the need to develop privacy-preserving genomic data processing and storage methods to benefit from cloud environments and keep private information secret.

Security attacks occur when an adversary has access or is able to modify data to which authorized access was not granted. Nowadays, DNA computing is used in biomedical sciences, with applications in healthcare and personalized medicine, where human DNA is used as a template, such as DNA molecules synthesis for abnormal cells detection. In this context, the synthesized DNA need to present properties similar to those of the DNA naturally found in the human cells. However, these similarities make the synthesized DNA susceptible to the privacy attacks reported on human DNA. Therefore, this section describes the security attacks on human DNA, which one should keep in mind when applying DNA computing for biomedical applications. When launching privacy attacks against human DNA, the adversary aims at discovering sensitive and not released information about a target individual or group. Privacy attacks were reported since 2006 and further exploitation of nonprotected genomic data was also described, with possibly severe consequences to the data owner, in particular, possible insurance denial and employment refusal [11, 12]. Due the demonstrated misuse of genomic data, researchers aimed at developing privacy-preserving techniques to adequately protect genomic data and prevent future harm for the data owners.

Fig. 1 presents a summary of all the different reported attacks on human DNA over the years for all the attacks categories. The genomic privacy attacks described in the literature can be classified into four categories, which are further discuss in this section:

1. Inference attacks
2. Re-identification attacks
3. Membership attacks
4. Recovery attacks

The main differences between the attack categories are the background knowledge (i.e., the information) the adversary has initially access to, and the kind of information the adversary tries to learn. Table 1 summarizes the adversary's background knowledge and target information for each category. The amount and quality of the background information directly impacts the outcome information of the attack, and consequently its success.

At the end, this section also discusses system exploits, which may target information systems that manipulate human DNA. Such exploits should be prevented so that privacy-preserving methods remain robust to attacks.

2.1 Inference attacks

Inference attacks aim at discovering additional sensitive information based on a partial or full genomic sequence. This kind of attack was also commonly used to infer the health status of target individuals based on their genomic sequence and on background knowledge about disease-related genes. The information retrieved from this kind of attack can be used by the other attacks categories. The background knowledge used for these attacks is some genomic information about a target individual (e.g., SNPs information, whole or partial genomic sequence) and population statistics such as allele frequencies or LD.

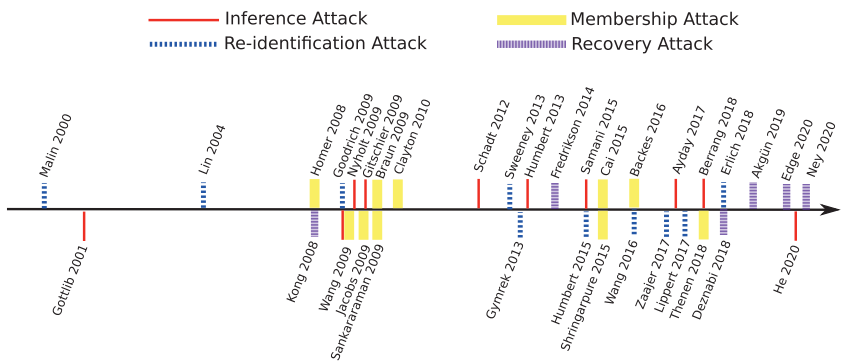

Fig. 1 Overview of privacy attacks.

Table 1 Security attacks: Background and discovered information.

Attack category	Possible background info.	Desired info.
Inference	Partial or full DNA sequence Population statistics	Hidden/nonobserved genomic variations Disease predisposition
Re-identification	Partial or full DNA sequence Available medical data Population statistics Demographic information Familial relationships Phenotypic traits GWAS statistics	Individual's identity
Membership	Partial or full DNA sequence Reference population statistics GWAS statistics Genotype frequencies Gene expression profiles	Participation of an individual in a group of interest
Recovery	Partial or full DNA sequence of relatives Familial relationships Genomic variation statistics	Reconstruction of partial or complete DNA sequence of a target individual

Gottlib [8] reported the use of genetic testing by US employers to infer the genetic disease susceptibilities of employees who did not provide their consent. Performing these tests without consent represents a privacy violation that can lead to employment discrimination. Nyholt *et al.* [13] demonstrated that even after hiding a particular gene, it might be possible to infer it from its neighboring regions. Despite Professor Watson's DNA sequence not containing information about the APOE gene—one of the main genes known to be related to Alzheimer's disease—the authors were able to discover its value using the neighboring regions of the hidden gene. This was made possible by the biological relations that link close regions in the human DNA. Wang *et al.* [14] proposed an inference attack using integer programming to infer hundreds of SNPs based on known pair-wise correlations between SNPs in the human genome. In this paper the authors also propose a membership attack that is discussed later in this chapter (see Section 2.3). Gitschier [15] described a method to infer haplotypes from the Y chromosome by exploring genealogical relations between men. The proposed method was used to infer surnames of individual from the Utah Residents with Northern and Western European Ancestry population

present in the 1000 Genomes Project. Therefore, it may also enable re-identification. Schadt *et al.* [16] performed an inference attack based on gene expression data, which allows the prediction of the genomic sequence that leads to an observed expression data. They then proposed the use of the predicted genomic sequence to perform a re-identification attack and discover individuals in large populations. Humbert *et al.* [7] demonstrated how to infer the genomic sequence of family members, related to an individual whose genomic sequence is observed. This attack uses reproduction statistics and known relationships between genomic variations in the human genome, and uses belief propagation, which in the context of genomic data is used to compute unobserved genomic variations that have a correlation with observed ones. Samani *et al.* [17] proposed an attack that takes advantage of publicly available high-order correlations among single nucleotide variations existing in the human genome, in particular recombination rate and diploid genotypes, to discover hidden or nondisclosed genomic variations. This attack has a higher inference power than previous work that considered only lower-order correlations. Ayday *et al.* [18] inferred hidden or nonobserved DNA information based on the partial DNA sequence of a target individual or a DNA sequence from a member of his family and publicly available phenotypic information. Berrang *et al.* [19] proposed a method that uses Bayesian networks, which leverage the combination of different types of information, for inference risk evaluation. The proposed method is used to perform inference of mother–child relationships based on DNA methylation profiles and genomic information, which the authors named a linking attack. He *et al.* [20] developed an inference attack based on publicly available genomic information and personal traits revealed by target individuals or their relatives. The attack allows an adversary to predict nonobserved genotypes and traits.

Table 2 summarizes the techniques used to perform inference attacks, the key findings, and the possible harm caused to the genomic data owner.

2.2 Re-identification attacks

Re-identification attacks aim at associating a given DNA sequence with its owner. The background knowledge that is required for performing such an attack might include publicly available DNA statistics, the genomic sequence of target individuals or publicly available genomic sequences, medical records with additional personal details (e.g., name, age, gender, geography), and genealogical databases that contain familiar relationships.

Table 2 Inference attacks.

Technique	Key finding	Outcome
Genetic testing	Genetic diseases susceptibility	Employment denial
Statistics-based inference	Hidden genomic regions inference	Disease susceptibility disclosure
Integer programming	Hidden genomic variations	Genomic information and disease susceptibility disclosure
Genealogy-based inference	Genomic profiles inference	Allow re-identification
Gene expression based inference	Genomic sequence inference	Allow re-identification and membership attacks
Belief propagation with genomic statistics	Hidden genomic variations	Genomic information and disease susceptibility disclosure
Genomic correlations-based inference	Hidden genomic variations	Genomic information and disease susceptibility disclosure
Phenotype-based inference	Hidden genomic variations	Genomic information and disease susceptibility disclosure
Bayesian networks	Mother–child relations	Familial relationships disclosure
Traits-based inference	Hidden genomic variations	Genomic information and personal traits disclosure

Malin and Sweeney [21] developed CleanGene, a software that assesses the identification risk associated to DNA sequences. This software performs re-identification based on available health-related data, i.e., from pharmacy records and hospital records, and disease knowledge from publicly available repositories. Later, the same authors proposed REIDIT [22], an algorithm that performs data re-identification. The proposed algorithm uses deterministic methods to link genomic data to named individuals, whose information is available in published records. The authors define a trail attack as a variant of a re-identification attack where the identity of a target individual is discovered using information that is collected from different independent sources (e.g., from different studies in which the individual participated). In the end, the adversary collects sufficient information to perform

re-identification, even if the information of each source alone is not sufficient. Li *et al.* [23] demonstrated that 75 statistically independent SNPs are sufficient to unequivocally identify an individual. The study was based on common random SNPs. However, if rare SNPs are observed the number of SNPs required for a successful re-identification would be even lower. This work provided an important baseline for further design of privacy-preserving approaches. Goodrich *et al.* [24] proposed the Mastermind attack, which demonstrates that even after applying cryptographic techniques to protect genomic sequences' privacy, it is still possible to learn some information at each guess attempt which goal is to learn hidden information (e.g., genomic variations). Like in the mastermind game, at each guess attempt, correctly guessed DNA positions are reported and if so, an adversary already learns part of the sequence. This attack allows the discovery of the identity of a genomic sequence owner with a limited number of guess attempts. Sweeney [25] proposed a re-identification method using demographic information applied to de-identified health data. This attack demonstrates that based on few attributes (i.e., place, gender, and birth data), which are combined with voter registration lists, it is possible to identify 53% of the American population. The information used in this attack was easy to obtain, since the voter registration lists was obtained for 20 dollars, and the health data was publicly or semipublicly available. Gymrek *et al.* [26] combined DNA information from the Y chromosome with publicly available genealogical information, in particular surnames. Then, taking advantage of the father to son surname heritage, the authors were able to de-anonymize 131 genomes from the 1000 Genomes Project. However, since this attack is based on the Y chromosome, it only applies to male individuals. Sweeney *et al.* [27] proposed an attack that links anonymized genomic data from the Personal Genome Project to their owner's name by combining publicly available records (i.e., voting lists) and demographic information (i.e., birth date, postal code, gender). Humbert *et al.* [28] performed a re-identification attack on anonymized publicly available genomic data that takes advantage of genotype–phenotype relationships, either obtained from public repositories that report SNPs and phenotypic trait relations or computed from genotype–phenotype databases. Wang *et al.* [29] demonstrated that it is possible to infer the identity of target individuals from aggregate statistics, such as GWASs statistics, even if they are differentially private, using Bayesian networks. Although the success of identity inference attacks decreases when rigorous privacy protection is applied to the GWAS statistics, the authors showed that the success probability of such attacks

increases with the background knowledge and that it is higher than that of random guesses. Lippert *et al.* [30] proposed a method that allows re-identification of individuals based on the prediction of traits applying phenotyping, and statistical modeling on Whole-Genome Sequencing (WGS) data. The authors relied on maximum entropy algorithm to use trait predictions for determining the genomic sample and phenotype profile that originated from the same individual. The authors show that phenotypic prediction from WGS data can enable re-identification without further data required. Zaaijer *et al.* [31] proposed a new method to perform re-identification of human DNA samples in a fast and inexpensive way, called MinION sketching. They demonstrated that analyzing 60–300 randomly selected SNPs and relying on Bayesian inference it is possible to link anonymized genomic samples with their owners. Erlich *et al.* [32] proposed a re-identification attack based on long-range familial searches. The proposed attack was performed on a dataset with 1.28 million individuals collected from direct-to-consumer services, such as 23andMe and Ancestry. Considering US individuals with a European ancestry, which represented 85% of the individuals from the dataset, it was possible to reach a third cousin or closer relationship in 60% of the searches. After finding a relative in a long-range familial search, the authors demonstrated that it is possible to perform re-identification using common demographic identifiers (e.g., age, gender, geography).

Table 3 summarizes the main techniques and background information used for re-identification attacks, and their key findings. In this kind of attacks the goal and harm caused is always identity disclosure by linking de-identified genomic data with the owner's identity, independently of the information and method used.

2.3 Membership attacks

Membership attacks focus on inferring the participation of a given individual in a study group. The attacks in this category compare the statistics of the target DNA statistics with the statistics of the target group and of a reference population. With this method, the closer the statistics of an individual are to a certain group, the higher is the probability that she belongs to that group.

Homer *et al.* [33] were the first to propose a membership attack that inferred the participation of an individual in a given study group. Using genomic variations expression or allele frequency information, the authors propose some statistics that compare the value obtained for a target

Table 3 Re-identification attacks.

Technique	Knowledge	Key finding
Medical data crossing	Pharmacy, hospital records, and diseases information	CleanGene software for re-identification risk assessment of DNA sequences
Deterministic methods to link DNA data to individuals	Public genetic and nominative information from different sources	REIDIT software performs re-identification; definition of trails attacks
Common allele-based statistics	Common SNPs information	75 Independent SNPs allow unequivocally identification of individuals
Data crossing	Demographic and de-identified health information	The identity of 53% of the American population was disclosed
Genealogical correlations	Y chromosome and genealogical information	Surnames disclosure for 131 genomes from the 1000 GP
Genotype–phenotype correlations	Genotype and phenotype information	Physical traits based re-identification
Bayesian networks	Aggregated statistics	Re-identification of a target individual participating in a GWAS
Phenotypic-based correlations	Whole-genome sequencing (WGS) data and physical traits	Re-identification without requiring further information
Long-range familial searches	Familial relations and demographic information	Familial relations reveal close relatives, and demographic identifiers allow re-identification

individual's DNA sequence with the ones obtained for a reference population and for a particular study group. Then, one can conclude that the target individual belongs to the study group, if its statistic are closer to the value of the study group. Interestingly, this attack showed that studying 50,000 independent SNPs is sufficient to infer the participation of an individual in a group of 100 people, or 10,000 SNPs if the group is composed of 10 people. Wang *et al.* [14] proposed an extension of the Homer's attack by adding into consideration existing statistical correlations among SNPs on the DNA, i.e., LD. Such correlations allow the inference of further nonobserved SNPs that

when included in the attacker knowledge empower the attack statistics. This work shows that the optimized attack is able to determine the participation of some individuals in a particular Genome-Wide Association Study (GWAS) even using a low-quality reference population. In addition, the attack used a couple hundred SNPs, which represent around 30× less SNPs than in Homer's attack to achieve the same attack power. Braun *et al.* [34] proposed the use of empirical tests to assess the participation of a target individual in a study group. The empirical test is performed using the target individual genotype information and the marginal allele frequencies of the group one is studying. Jacobs *et al.* [35] demonstrated how likelihood-based statistics can be used to infer the participation of a target individual or his close relatives in a GWAS. The statistics are computed using genotype frequencies and individual genotypes. In addition, this paper evaluates the membership attack power for different sample size and considering different sets of SNPs for computing the statistics. Sankararaman *et al.* [36] developed the SecureGenome tool, which enables the detection of a target individual on a study group based on the summary statistics from a GWAS. The membership attack compares the target individual alleles profile with the allele frequencies of the study group and the allele frequencies of a reference population, similar to Homer's attack [33]. Clayton *et al.* [37] designed a membership attack using a Bayesian approach. This attack considers prior probability knowledge about the participation of an individual in a certain sample. Shringarpure and Bustamante [38] demonstrated an attack on the Beacons Project. This project designed a platform for secure querying of genomic information where only Boolean answers are returned to the user, with the purpose of limiting the amount of private information disclosed. Although the information obtained per query is limited, the proposed attack shows that by querying 250 genomic variations and combining their results, it is still possible to discover the participation of an individual in a beacon with 64 European individuals. Re-identification was also deemed possible; however, it required a much higher number of queries (1000 genomic variation queries). This attack showed that beacons are susceptible to membership attacks and may also leak phenotypic information about the participants they study. Cai *et al.* [39] proposed an attack that takes as input a dataset and the GWAS statistics from 25 randomly selected genomic variation sites to infer whether an individual participated in the case group. More precisely, this attack requires the genotype of a target individual and compares it with the genotypes inferred from the case group using the GWAS statistics, and if a match is found, the target individual is

considered to have participated in the case group. From then on, it is also possible to perform re-identification of case individuals. Backes *et al.* [40] proved that genomic data is not the only type of omics data that can be used to perform membership attacks. They used expression data, in particular from microRNA and showed how to infer the participation of a target individual in a particular group. Since microRNA data is more affected by the health status than genomic data, it is therefore more informative about the group the individual belongs to, be it the control or the case group. Thenen *et al.* [41] improved the membership attacks proposed by Shringarpure and Bustamante. With only 5 queries (50× less queries than the ones required by the former attack), they were able to infer the participation of individuals in a beacon with 95% of confidence, using the same beacon configuration used in previous works. The attack improvements are due to the use of high-order Markov chains to infer high-order relations on the genomic data. Another important finding of this work is that current privacy protection measures, which include particular genomic regions hiding and the implementation of a query budget, are not efficient against the proposed membership attacks.

Table 4 summarizes the techniques and background information used in the membership attacks, and their key findings. All the attacks in this category require some genomic information of the target individual(s).

2.4 Recovery attacks

Recovery attacks focus on inferring the DNA sequence of a target individual aided by publicly available DNA statistics, such as allele frequency in the reference population, and/or kin relations if the genomic sequences of relatives are available. The inferred DNA sequence can then be used to perform attacks from the previous categories, which assume the availability of the DNA sequence of a target individual.

Kong *et al.* [42] exploited kin relationships to infer haplotypes of target individuals based on observable genomic information of their relatives. The proposed inference method incorporates information about recurrent mutations transmitted from parents to their children and fine-scale recombinations. Wang *et al.* [14] also performed recovery of nonobserved sequences, overall 100 sequences containing a total of 174 SNPs were recovered, based on single and pair-wise allele frequencies. Those recovered sequences were then used in the proposed inference and membership attacks. Fredrikson *et al.* [43] showed that personalized medicine models

Table 4 Membership attacks.

Technique	Information	Key finding
Queries knowledge combination	Genomic variations expression data or statistics	50,000 SNPs are sufficient to disclose membership in a group of 100 people and 10,000 SNPs for a 10 people group
Statistics comparison	Genomic variations expression data or statistics and statistical correlations	Including statistical correlation in the genome requires 30× less SNPs to achieve the same attack power
Empirical tests	Allele frequencies of the study group	Membership disclosure
Likelihood-based statistics	GWAS statistics	Membership inference of a target individual and close relatives
Bayesian methods	Target group statistics	Membership disclosure
Queries knowledge combination	Genomic variations queries	For a beacon with 64 European individuals 250 queries allow membership disclosure
Higher-order Markov chains	Genomic variations information	Membership prediction using 50× less queries for the same result of similar attacks
Disease and gene expression correlations	Expression data	Membership disclosure in disease-related studies

can leak information about an individual's DNA sequence. By combining the information from a pharmacogenetic model, which was used to design particular medicines for a patient, and demographic information from the same patient, the authors proved that is possible to discover some hidden regions of the DNA sequence of a patient. Deznabi *et al.* [44] described how to discover parts of the genomic sequence of a target individual based on familial relations, public phenotype information, and other available data from online repositories (e.g., social networks). Akgün [45] proposed an active recovery attack which allows the adversary to discover genomic data from an individual using SNP statistics. This attack consists of the manipulation of the weights attributed to the SNPs used in a test so that it is easier to infer the SNPs of a target individual from the test results. This attack was one of the few assuming a dishonest party. Edge and Coop [46] proved that using

publicly available genomic data an adversary is able to learn the genomic sequence of a target individual using Identical-By-State (IBS) tiling. This technique consists in matching known genomic sequences against an unknown one to obtain information about it. The authors showed that applying IBS tiling for 900 genomes from the 1000 Genomes Project reveals at least one allele from 82% of the SNP sites of an individual with European ancestry. In addition, the authors proposed a variant of IBS tiling, called IBS probing, that allows the adversary to learn if the target individual' genome contains a specific disease-related allele, whose neighboring sequence is known. A related attack was described by Ney *et al.* [47]. In this attack, the authors demonstrated that an adversary could almost learn the entire genomic sequence of a target individual from GEDmatch—a US direct-to-consumer online service that compares DNA data files. Two possible ways are described to learn the individual's genomic sequence: (i) by uploading artificial nearly-all-heterozygote genome and examining the resulting IBS segments (similar to Ref. [46]), and (ii) by uploading an all-heterozygote genome and examining the resulting images.

Table 5 summarizes the different techniques and background information used on recovery attacks, and their respective key findings.

2.5 System exploits

System exploits can affect all information systems, and are therefore not specific to genomic data processing systems. They explore system vulnerabilities and they must be taken into account when designing privacy-preserving systems in order to ensure their long-term security. System exploits and intrusions can lead to user's data exposure and consequently to information leakage. In order to be secure, systems must ensure confidentiality, integrity, and authentication. Confidentiality focuses on preventing unauthorized access to the data. Integrity ensures that the data is not modified by unauthorized users and is also in charge of reporting those changes in case they happen.

Table 5 Recovery attacks.

Technique	Information	Key finding
Kin-based inference	Kin relationships and relatives partial genomic data	Target individual sequence reconstruction
Information integration	Pharmacogenetic models and demographic information	Target individual genomic sequence reconstruction
Identical-by-state tiling	Known and unknown genomic sequences	Multiple SNPs inference that allow genomic sequence reconstruction

Last but not the least, authentication is the property that allows the verification of users' identity and then grant or deny them access to the system.

Malicious attacks are also part of system exploits that can lead to privacy breaches and data theft. A successful attack can result in data loss if backup data copies are not maintained.

Finally, in the context of DNA data, maintaining secure systems is of paramount importance, due to the previously discussed sensitive information encoded in the DNA and the reported privacy breaches. There are two main scenarios for genomic data systems: (i) the data is stored locally or in a private server, or (ii) the data is stored in a public cloud, e.g., because of its large size.

In the first scenario, the system designer is responsible for placing protection methods to prevent system exploits. To strengthen the protection, protection techniques can be also applied at the data level, for example, data obfuscation and data encryption. While, for the second scenario, the system protection is of the responsibility of the CSP. Therefore, the user should protect his data before sending it to the public cloud. Commonly, this process is made through data encryption since it prevents the cloud or any other entity that obtains an access to the data to learn its real value.

3. Privacy-preserving techniques

Traditional privacy-preserving techniques need to be adapted to be used on genomic data, since, as discussed previously, genomic data itself contains re-identifiable information. Several approaches were proposed to perform some computations on genomic data in a privacy-preserving way, such as in GWASs statistics computation, DNA sequences alignment, and genomic database queries. GWASs consist in the analysis of several genomes over multiple genomic variation positions to find the relation between genotypes and diseases. Such solutions use different techniques, such as data obfuscation, cryptography, and trusted hardware. DNA computing itself is a paradigm that can be used to design novel privacy-preserving techniques; however, this field is in its infancy. Briefly, work on this field described DNA cryptography, which is described later in this section.

As the previous section detailed, several privacy attacks on genomic data have been described in the literature. Following these findings, the potential impact of privacy attacks on data owners made the research community focus its effort on the development of methods to prevent successful attacks. These methods can be categorized as follows:

1. De-identification methods
2. Data augmentation methods

3. Cryptography-based approaches
4. Secure Multiparty Computations (SMPC)

3.1 De-identification methods

De-identification consists in removing all the personal identifiers from the data in order to keep secret the identity of the data owner. The main goal of de-identification is to prevent the direct association of genomic data with their owners, and consequently, protect the owners identity. In other words, de-identification aims at providing data anonymity. This is a technique widely implemented for biomedical data. However, as demonstrated by re-identification attacks, applied alone this method is often not enough to protect the data owner. Indeed, these methods are particularly inefficient in protecting genomic data since they do not remove the identifying information contained in the genomic data itself (i.e., rare genomic variations). In addition, providing anonymity has become more difficult with greater availability of information in online platforms, which can be related with genomic data and contribute to individuals' identification.

K-anonymity is a widely applied paradigm to enforce a stronger data de-identification, which consist in modulating the data attributes in such a way that based on those attributes an adversary is not able to distinguish an individual from k-1 other individuals [48, 49]. The two methods mainly used to achieve k-anonymity are suppression and generalization. Suppression consists in removing the attributes that can lead to direct identification of an individual, such as those that are not shared with other individuals in the dataset and are not generalizable. Generalization consists in translating an attribute value in a broader class. For example, it is common to replace a numerical value by an interval that contains it. Emam *et al.* [50] developed a de-identification algorithm that ensures k-anonymity on health datasets. Other commonly used methods to achieve anonymity are l-diversity and t-closeness. l-Diversity [51] was proposed as an improvement of k-anonymity that preserves privacy when the diversity in the attribute values is low, and assumes that the adversary has access to background knowledge. t-closeness [52] is a refinement of l-diversity where the difference between the distribution of the sensitive attributes for a given class and the distribution of the same attributes in the full table is at most t. Although this method improves privacy, it also implies some utility loss at the data management and data mining levels.

DNA Lattice Anonymization (DNALA) [53] was proposed to anonymize pairs of genomic sequences, which are represented by the sequence that represents the minimal distance to both. This is a generalization process that resembles k-anonymity. The proposed method was tested on human genomic sequences publicly available. The main limitations of the proposed generalization are the following: (i) it is dependent of the pair of sequences considered; and (ii) it is limited to two sequences.

Lin *et al.* [54] proposed a generalization method based on data binning. The proposed approach ensures that no unique record is present in the database released to the users. The bin size works as an anonymity level indicator, since a larger bin size makes data less specific and detailed, and consequently, provides a higher level of anonymity.

In conclusion, although de-identification methods were reported to be insufficient to prevent re-identification [38, 55, 56], they do complicate them. It is also important to consider that anonymity in the context of genomic data is different from other data types, since genomic data contains personal identifiable information itself. Furthermore, its combination with other metadata, such as name, gender, age, and geographic details, power the privacy attacks described in the previous section. DNA computing, as an emerging paradigm, could also contribute for the development of de-identification methods, allowing the data to be de-identified as soon as it is produced by the sequencing machines. This could be done, for example, by removing the need of metadata, such as identifiers, by also encoding them in a DNA sequence format.

Fig. 2 summarizes common de-identification techniques used for medical records. For the name and diagnosis columns, it is used pseudo-anonymization where the real names and diagnosed disease are replaced

Name	Age	Num. exams	Diagnosis
Alice	34	3	Diabetes
Bob	25	10	Cancer
Claire	29	7	Diabetes
David	31	4	Diabetes
Eva	28	2	Cancer
Frank	28	9	Cancer

Name	Age	Num. exams	Diagnosis
P1	30-34	1-5	D1
P2	25-29	6-10	D2
P3	25-29	6-10	D1
P4	30-34	1-5	D1
P5	25-29	1-5	D2
P6	25-29	6-10	D2

Fig. 2 De-identification.

by a unique identifier. For the age we have two classes (25–29 and 30–34 years) that grant 4-anonymity for the 25–29 class and 2-anonymity for the 30–34 class. In other works, this means that at least four records have the same information in a given data column. For the number of exams, there are two classes (1–5 and 6–10 exams) and this generalization grants 3-anonymity.

The referred techniques are applied to metadata that usually is together with the DNA data; however, DNA data contains identifiable information itself. Therefore, other techniques are required to completely anonymize the DNA data.

3.2 Data augmentation methods

Data augmentation consists in applying generalization or obfuscation in order to protect data. In the context of genomic data, it consists in making the data of different individuals indistinguishable by generalizing or obfuscating the information it contains, so as to prevent their unequivocal identification.

Generalization, at the genomic sequence level, consists in the representation of two or more sequences with the most common sequence among them. In other words, a set of genomic reads are represented by the most common genomic variations they contain [53, 57].

Data masking consists in hiding sections of the data in order to make them unobservable and unpredictable for an adversary. The data that is masked corresponds to the sensitive information one wants to protect from unauthorized access. Cogo *et al.* [58] proposed the first automated sensitive DNA short sequences detection, which relies on Bloom filters. This approach improves privacy protection by allowing the user to store and process the sensitive and insensitive DNA sequences differently. Later, Decouchant *et al.* [59] proposed an automated sensitive information detection for long DNA sequences. This approach allows the efficient privacy-preserving processing of DNA sequences with a lower performance overhead and higher precision than Cogo's approach. Extending this approach, Fernandes *et al.* [60] designed a sensitivity levels classification method-based DNA properties, such as allele frequency and LD.

Differential privacy is another data augmentation technique, which is used to make an aggregate result indistinguishable whether a single individual participates or not in that result through the addition of noise [1, 61, 62]. With differential privacy, the greater the noise added, the higher the privacy

protection is. However, the addition of noise reduces the data utility. Consequently, applying this technique requires studying the trade-off between data utility and privacy protection to ensure that subsequent data analysis will not be compromised.

The techniques in this category make re-identification attacks harder to perform, since their main purpose is to hide sensitive information or genomic regions, possibly by adding noise.

3.3 Cryptography-based approaches

Cryptography-based approaches are characterized by the protection of the input and output using an encryption scheme. These approaches can be mainly used for two purposes: storage or processing. They differ by the encryption schemes used since the processing scenario requires operations to be allowed on the encrypted data. Cryptography-based approaches are interesting because of their guaranteed privacy-protection; however, their application is limited due to their longer computational time. Garbled circuits [63] allow two parties to perform secure computations. For example, a user can send his encrypted data to a server where some computations are performed and, then, the encrypted results is sent back to the user. Garbled circuits protect the input and intermediate results, which are never revealed, since they are always manipulated encrypted on the server side.

Homomorphic encryption is a particular subject of cryptography which allows mathematical operations on the encrypted data. This allows some computations to be outsourced to untrusted environments, such as public clouds, without having to decrypt the data. Atallah *et al.* [64] designed a privacy-preserving strings comparison algorithm using homomorphic encryption, which computes the edit distance between two DNA sequences. Kantarcioglu *et al.* [65] proposed a cryptography-based approach that allows genomic sequences sharing and querying.

Despite the performance limitations of cryptography-based approaches, there are still some applications where encryption can be practical, such as determining disease susceptibility through genetic testing [66] and secure datasets querying [67]. He *et al.* [68] proposed a cryptography-based approach to identify relatives. In this approach a pair of individuals only share the necessary encrypted genomic information to determine if they are relatives or not. The results obtained showed that this approach is able to find relationships up to third cousins level while preserving the privacy of the individuals.

Cryptography-based approaches provides a high level of protection; however, current encryption schemes are not designed to protect genomic data for its full lifetime [7]. This is a real challenge since an individual's genome is partially inherited from previous generations and transmitted to the future ones.

In the context of DNA computing, DNA cryptography is emerging. This technology uses DNA algorithms that convert the information to be protected into DNA, first converted to the digital DNA sequence and then to the DNA molecule. This process is equivalent to the operations performed by standard cryptographic schemes, where plaintext information is converted into encrypted data. In this field two types of DNA cryptography methods exist: DNA-based data hiding schemes and DNA-based encryption schemes. DNA-based data hiding schemes consist in converting the message to DNA nucleotides and then mix the real message DNA with fake DNA sequences and send the mix to a receiver. DNA-based data hiding schemes consist in encoding the information in DNA nucleotides instead of using binary form. Then, the double helix is created following the complementarity property to increase the complexity [69, 70].

Fig. 3 summarizes the process of encryption of DNA sequences. Usually, the DNA sequence is translated into a 2-bit sequence and, then, each 2-bit is encrypted individually as represented in the figure. Following, computations can be performed on the encrypted data and in the end the encrypted result is decrypted to reveal the result in clear text.

3.4 Secure multiparty computations

SMPC allow the secure collaboration of several institutions, e.g., hospitals, biocenters, researcher centers, and universities. SMPC enable computations

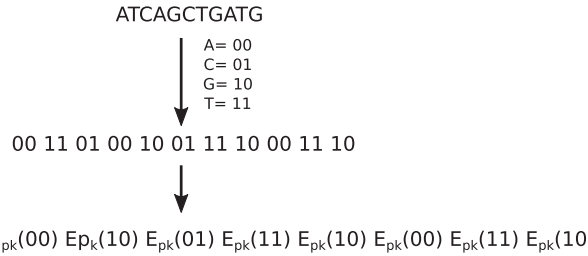

Fig. 3 DNA encryption.

over a dataset that is distributed among the participating institutions while keeping the data private [71]. This technique allows third parties to perform computations on encrypted data, without learning any information about the data, the computed results, or the contribution of any party participating in the computation. Therefore, SMPC does not require to trust a third party. In addition, since data is transferred encrypted, data usability is not compromised while providing privacy protection. On the other hand, the computational overhead and communication costs between participants are high, which may make SMPC not practical for some applications. SMPC can use standard cryptographic algorithms, such as AES, which provide strong privacy and confidentiality properties. Some examples of SMPC-based approaches include the work described in Refs. [72–74]. Aziz *et al.* [72] proposed a SMPC-based approach that used the Paillier encryption scheme to enable queries on genomic data. The results were obtained considering five geodistributed parties and show that this method requires a computational time comprised between 7 min and 4 h. Cho *et al.* [73] proposed secure GWAS analysis through SMPC. Another closely related contribution is METIS [74], which uses SMPC to make the clients, the server, and data owners cooperate to enable computations on genomic data provided by the data owner without disclosing it to the other parties. This approach intents to give the genomic data owner the full control regarding his/her genomic data analysis permissions. At the same time, the data owner has no computations overhead.

Table 6 summarizes the advantages and disadvantages of the privacy-preserving techniques discussed in this section.

Table 6 Overview of privacy-preserving techniques.

Technique	Advantages	Disadvantages
De-identification	Personal identifiers removal Replace identifiers by pseudo information.	Not efficient. Genomic data is identifying information itself Permit to link genomic data to owner's identity
Data augmentation	Hidden information Real value replaced by more abstract information	Data utility can decrease Information precision decreases

Continued

Table 6 Overview of privacy-preserving techniques.—cont'd

Technique	Advantages	Disadvantages
Cryptography	Highest protection Prevent unauthorized access to the data Computations secure computations running on untrusted environments	Low performance Data size increases Computational overhead increases Limited operations allowed
SMPC[a]	Allow secure collaboration Computations run on encrypted data Data utility preserved No computational overhead on the client side	Large data analysis require nonnegligible data transfers High computational overhead

[a]Secure multiparty computations (SMPC).

4. Trust management

Trust management considers other possible sources of exploits that can compromise the genomic data privacy. The data producer, the users, and the storage service providers play an important role on the genomic data cycle. Privacy protection for genomic data must consider the vulnerabilities to potential attacks on the data producer side, on the user side, and on the storage provider. This topics are discussed in the following sections.

4.1 Genomic data ownership

An important question regarding human DNA privacy and trust management is to determine to whom the sequenced genomic data belongs to, and who should keep it. For several researchers the sequenced genomic data belong to its donor and he/she should own the full rights regarding that data. However, nowadays, DNA sequencing and processing involves different institutions (i.e., sequencing center, biocenter, hospital) which make ownership complex with several entities requesting and having access to the DNA sequence. As discussed in the next section, the data provider is a commonly trusted entity, whose main role is to produce the DNA sequence. However, assuming that the DNA sequence should be kept only by its donor, the traditional processing pipeline needs to be modified so that

the data provider or any other entity involved in the data processing is not able to keep a copy of the data.

In the processing phase, genomic data owner should be able to define for which applications he/she allows the data to be used, in other words, the data owner must define the data access and usage policies. There are two main kinds of data use, primary and secondary. The primary use of the data grants the permission to use the data for a particular application (e.g., research, GWAS, genetic testing), which is often the original reason behind its collection. The secondary use allows the data to be used for other applications than the initial use case. In both cases, the data is controlled by the service provider. Therefore, the data owner needs to define the access and usage policies before the access to the data is granted. Furthermore, at any moment the data owner must be able to update the consent accordingly to his/her preferences, which should be immediately implemented. Although, the data owner might allow the use of his/her data and later change it, once the data usage has been granted for some application, it is impossible to ensure that the involved entities destroy their data copy.

4.2 Trusting the data provider

One way to get a genome sequenced is to participate in genetic studies by donating a biological sample, which is handled by the data provider. The data provider, commonly biocenters with sequencing facilities, is the entity responsible to generate the data. The data providers for human sequencing data is the sequencing center. In the DNA computing context, the data provider is the laboratory where the DNA sample is synthesized.

Other widely known data providers are direct-to-consumer sequencing services, such as 23andMe, Helix, MyHeritage, and Ancestry, that provide genome sequencing service and allow the user to explore its own genomic information as a recreational process. The cost of those services has been dropping in the last decade (23andMe: from $99; MyHeritage: 79€) and their use has increased. However, when using such a service the user must inform himself about the policies of data ownership. It must be clarified by the service provider if the spare biological sample is discarded and whether the genomic data is completely deleted upon request. Nevertheless, most of these direct-to-consumer sequencing services store the genomic data in servers they rent from public clouds, where it can be accessed by the user. Therefore, the user must also get informed about the data protection policies while choosing a service provider.

One other point of discussion is whether or not the data provider should keep a copy of the data. From the privacy point of view, it would open a window of vulnerability if a copy is kept, since the data is owned and managed by another entity. On the other hand, having a copy of the data allows better availability and reduces the risk of data loss. This decision must take into account the specific data storage and sharing policies of each data provider.

4.3 Access control

Access control consists in regulating who has access to the data following stipulated access policies and agreements. In addition, access control is also in charge of ensuring the traceability of the access to a particular file or data. In other words, it can be described as registering all the accesses made to the data, which in case of data modification or corruption would allow the identification of its author [62, 75]. Although this technique alone is not privacy-preserving, since it does not itself provide data protection, it is widely used in combination with other techniques.

Nowadays, the most commonly used access control model is the role based access control (RBAC) method, in which access is granted to the users that have a role that justifies data access for specific tasks they are in charge of. In the context of DNA data, the access to the data would be granted to the specialists the owner accepts to share data with, e.g., doctors and researchers.

Several genomic sequence repositories require a data access request that is evaluated by a specialized committee. Although this process is essential to protect the data from being misused, it can take several months until it concludes. This process can even be longer in the case of collaborative repositories since all the involved institutions need to approve the access. Another limitation of this methodology is that the access to the data is commonly granted for a short period of time, and in case the access period needs to be extended, the decision process needs to be repeated [76]. This process can sometimes limit accessibility to the datasets in exchange for more control over data accesses and, consequently, reduce the possibility of privacy breaches.

Erlich *et al.* [77] described an alternative to the classic access control, bilateral consent framework (BCF), to facilitate genomic data sharing while protecting data privacy. In the BCF, the participants can directly decide who can access their data instead of requiring a decision from the access control committee. This streamlines the process of obtaining access to data and also

to apply changes on the access rights. However, although such a model seems to be promising, the privacy sensitive nature of genomic data makes many institutions keep utilizing more classic access control schemes. This demonstrates that dynamic access control solutions require further development.

Blockchain is an emerging technology that allows the collaboration between multiple parties removing the need of an intermediary trusted centralized party to authenticate the interactions. In the context of healthcare, blockchain applications have been increasing, with particular interest for patients identity's validation, to facilitate the management of permissions, and for access control to biomedical data [78, 79].

In the DNA computing field, Namasudra proposed a secure and fast access control model to address the system overhead and the long accessing time required when searching for data stored in the cloud [80]. This scheme keeps a fast data access list on the cloud side, and uses a 1024-bit DNA computing based key for data encryption, which increases data security.

4.4 Cloud environment: Storage and processing

Genomic data processing requires extensive resources due to the huge amount of data to be analyzed and the intensive computational processing. Therefore, this processing is often outsourced to public clouds, which provide performance improvement and larger storage memory. However, clouds are managed by a CSP that provides limited control to the user over its own data. Furthermore, the data can be manipulated (i.e., copied, transferred) by the CSP without users awareness and stored in multiple locations for data availability. These properties make the data more prone to possible unauthorized access by the CSP or an intruder when it is stored in a public cloud [81]. In order to prevent the unauthorized access to genomic data on clouds, the user can apply some protective techniques and must agree with the CSP on the conditions to handle the data [82].

In particular for genomic data, it is also important to know the data properties to limit the potential information leakage in case the data outsourced to the cloud might suffer some privacy breach. In Ref. [83] the authors define some data aggregation properties, in particular the ratio between number of genomes and number of genomic variations that users should follow when releasing genomic data to prevent information leakage. Cryptography-based schemes have been proposed to allow the outsourcing of intensive computations, such as sequence comparison, to public clouds while protecting data

privacy [84]. Chen *et al.* [85] proposed a privacy-preserving reads alignment approach that combines processing on public and private clouds. The main concept of this approach is to assign hash values to sensitive information and, then, process the hashes in a public cloud. Further processing on the sensitive information is performed in the private cloud. Although this process outsources part of the computations, significant computations still run in the private cloud. Balaur [86] performs alignment on hybrid clouds based on MinHash and k-mer voting. This approach was developed to allow the user to transfer part of the processing to the public cloud while guaranteeing privacy protection and ensuring high accuracy. The candidate positions step uses privacy sensitive information, i.e., the genomic sequence and the reference genome in plaintext, and is performed in the trusted environment (private cloud), while the secure voting that selects the best position a read aligns to is performed using hash values in the public cloud. DepSky [87] focused on providing secure storage in a cloud-of-clouds relying on encryption, encoding, and data replication. Charon [88] was designed to provide secure storing and sharing of genomic data on cloud-of-cloud systems. GenoShare [89] is a tool that was developed to support conscious genomic data sharing. This tool takes as input the genomic information to be shared, the adversary's background knowledge, and other publicly available information, for example, previous data releases. It then simulates three privacy attacks (membership, phenotype inference, and kinship inference) to determine which data can safely be shared and which data should not be shared. More recently, Cogo *et al.* [90] proposed a privacy-preserving efficient and dependable cloud-based storage approach for human genomes. This approach combines sensitive information detection and deduplication. The sensitive information detection method is applied to human genomic reads for privacy protection improvement, while the deduplication method allows the optimization of the storage space. This work represents an important effort toward privacy-preserving outsourcing of genomic data to the cloud environment.

Trusted Execution Environments (TEEs) are secure and isolated memory and code partitions located on a processor, which are not accessible to the remaining system. They were developed to provide data confidentiality while ensuring the integrity of the code they execute. Consequently, if communications with the TEEs are secured, the data and intermediary results are protected. Some examples of TEEs include ARM TrustZone and Intel SGX. However, some side channel and cache attacks were

reported and showed vulnerabilities of TEEs. In reaction, mitigation methods have also been proposed [91, 92]. Some secure genomic data processing algorithms have been developed using TEEs, including secure genetic testing and privacy-preserving decentralized processing of genomic data [93–96]. However, to assume that the processing done inside a TEE is secure, it requires to trust the processor manufacturer, e.g., Intel for the SGX technology.

Privacy risks associated to the processing of biomedical data in cloud environments were studied in Ref. [97]. This study highlights the main benefits of cloud computing as the following: (i) vast resources availability, which is useful for parallelization; (ii) affordable cost; and (iii) the user is free from the maintenance duties. However, important security challenges also appear when outsourcing sensitive information to the cloud, which include: (i) control, management, and security of the data is the user responsibility and the CSP; therefore, they should agree on the policies; (ii) access rights need to be well defined and updated, since cloud environments can be accessed remotely; (iii) the user must prevent data loss by holding a backup copy, since cloud environments are shared resources and are susceptible to failures, as all computational systems.

Table 7 summarizes the discussed trust management fronts in this section, highlighting the main challenges and techniques used.

Table 7 Trust management.

Topic	Challenges	Techniques
Data ownership	Define who owns the data	Access and usage agreement
Data provider	Ensure the data is used for the only purpose it was created Guarantee that no copy is used without permission.	Informed consent when sequencing
Access control	Prevent unauthorized access to the data Provide access traceability	Blockchain Data access lists Bilateral consent framework
Cloud environment	Protect data from being accessed by intruders and the Cloud Service Provider (CSP) Agreement between CSP and user regarding data storage and access	Data encryption Trusted execution environments Hash functions

5. Discussion

Genomic data presents a wide range of applications, which boosted its production and availability, creating massive amounts of genomic data to be processed. This led the research community to develop high-throughput algorithms and search how to leverage more efficient environments for genomic data processing, such as public clouds. However, several privacy attacks on genomic data have been reported in the literature, demonstrating that data was not being protected adequately. Such privacy attacks are classified into four main classes: inference attacks, re-identification attacks, membership attacks, and recovery attacks. The main difference between the different types of attacks is the information the adversary intents to discover, ranging from obtaining additional genomic information about a target individual to discovering the data owner identity. This applies to human DNA data, which is characterized by natural correlations that occur between regions in the genome and between the genomes of members of the same family. Privacy attacks on human DNA use those correlations to improve the power of the attack. For synthesized DNA such correlations exist if it was created based on the human genome. If it is created independently, the properties can be different; however, since they share the same format, similar attacks might be devised to disclose information contained in synthesized DNA. The potential harm caused by privacy attacks alerted the researchers about the urgent need for efficient privacy-preserving approaches for storing, sharing, and analyzing it. Although several privacy-preserving approaches were proposed in the last decades, the field is still at an early development stage and reaching a good balance between privacy and utility is still a challenge in many applications. The main difficulty to address is that privacy-serving techniques commonly used are not sufficient for genomic sequences since they contain personal and identifiable information themselves.

Privacy-preserving techniques applied to genomic data include de-identification methods, data augmentation methods, cryptography-based approaches, and SMPC. In the de-identification methods, anonymization techniques are widely used for genomic data; however, they have been shown to be insufficient to prevent re-identification attacks [26]. For genomic data, anonymization is not sufficient since the data itself contains personal identifiable information. Nonetheless, those techniques are still applied to genomic data jointly with other protection methods, since they make the identity inference task harder. Data augmentation methods are

more efficient than anonymization since they aim at hiding or generalizing properties that are inherent to genomic data. Data masking and differential privacy are commonly applied methods. Such methods target the prevention of re-identification attacks, even though they might not completely prevent them. A drawback data augmentation methods is reaching an acceptable trade-off between privacy and utility loss. In particular for differential privacy, in order to achieve good privacy protection, considerable noise is added to the data; however, its utility might become compromised. Cryptography-based approaches provide the highest protection to genomic data and they are very efficient against all kinds of privacy attacks, since any data or metadata is not revealed to any unauthorized party. However, the required computational resources and time are not negligible and even unpractical for some processing tasks. Cryptography-based approaches are efficient to prevent all the described privacy attacks if the data is only handled in plaintext in trusted environment and by the authorized users. SMPC are often applied to allow collaboration between different entities, e.g., biocenters, that want to perform computations over all data parties without revealing each entity data share. Regarding privacy protection, SMPC is efficient since data is only transmitted encrypted and the computations are only decrypted by the legit users, which have the decryption key.

Giving a practical example, to prevent recovery attacks as described in Refs. [46, 47] that apply IBS tilling and combine publicly available genomic data and genealogical databases, two mitigation approaches could be implemented. First, often the IBS tiling reveals the location of the queried segments, which contribute to the inference of the target individual's genome. Therefore, hiding this information from the results would complicate the sequence inference since the adversary would only learn if the segment is present or not in the genome. Second, for the case of the use of artificial or manipulated genomic sequences, the prohibition of such sequences would prevent the techniques applied in Ref. [47]. However, authentication techniques for genomic sequences are still an open challenge.

For synthesized DNA, natural correlations occurring in the human DNA are not considered, which introduces some independence between synthesized DNA sequences. Thus, techniques such as data augmentation and cryptography-based techniques can be applied and might provide improved protection to synthesized DNA. To determine the information encoded on the synthetic DNA molecules, traditional sequencing technologies that have been designed for human DNA sequencing can be used. This process eventually returns the plaintext sequence of nucleotides corresponding to the

synthetic DNA information, which in turn reveals information about the original data that was used to create it. Therefore, the sequenced data must be stored securely to prevent information leakage.

The described privacy-preserving techniques are yet insufficient to address the performance and utility that would fully take advantage of genomic data processing while ensuring the adequate privacy protection. Interoperability among genomic data from different institutions potentially geodistributed is still not fully addressed. In addition, in this scenario standardization and common laws are required to replace the existing per country genomic data privacy guidelines.

In the context of genomic data, trust managements need to consider several points. First, the data owner must have the right to manage the data, including data access and processing. However, there are open discussions regarding who should keep the data, with the following possible scenarios: (i) only the data owner has a copy of the data; (ii) the data provider has a copy of the data; and (iii) the entities involved in the processing have a copy of the data. In scenario (i), the main problems are that data availability and data protections are solely of the responsibility of the data owner. In this case, data access would need to be requested to the owner each time the data is requested. In addition, there is a higher risk of data loss and in case it happens the genome needs to be re-sequenced which incurs extra expenses and extra time. In scenario (ii), the data owner needs to allow the data provider to keep a copy of his/her data. This scenario might leave the genomic data more exposed than in scenario (i), since the spare copy is managed by the data provider and the data owner needs to trust it. However, in case the data owner loses his copy, it can ask the data provider a new copy. In scenario (iii), allowing the processing entities to also have a copy of the data, means even less control on the data accesses. Although currently access control is strictly managed, involving other entities increases the complexity of the process. On the other side, this scenario provides better data availability.

Although access control have been defined as essential on systems storing and processing genomic data, its implementation is not yet optimized. Nowadays, data access requests are reviewed by a specialized committee and last several months.

The storage and processing of genomic data are often outsourced to more cost-efficient and powerful environments, such as public clouds. However, processing biomedical data on public clouds can represent an increased information leakage risk when the data is not efficiently protected.

Those risks have been described in Refs. [5, 6]. Since the public is an environment accessible to multiple entities, the data stored in it is more prone to possible access attempts by unauthorized entities. In order to take advantage of the computational resources provided by cloud environments, several solutions have been proposed to perform privacy-preserving computations on genomic data. However, such solutions are limited to few possible computations and they require nonnegligible computation on a trusted environment. Therefore, further studies are required on this subject in order to allow the maximization of the resources without sacrificing performance or privacy.

To conclude, privacy-preserving processing of genomic data has been evolving in the last decades. Significant advances were made; however, the journey is still not over. In addition to the improvement points discussed in this section, some other open challenges include: (i) considering a more realistic threat model, which would include malicious adversaries; (ii) considering dynamic systems, which tolerate datasets modifications (Refs. [98, 99] are early examples for genome wide association studies and genomic beacons, respectively); (iii) ensuring interoperability between data from different sources and define genomic data privacy standards. First, the most common threat model considered for the privacy-preserving approaches design is the honest-but-curious adversary. This model describes an adversary that follows the protocol but might however try to learn further information about the data. The honest-but-curious adversary model is a moderate model, since the behavior of the adversary is somehow controlled. In practice, malicious adversaries, which are adversaries that intentionally try to deviate from the protocol to extract information they do not have the right to access, or perturb the system, form a more realistic threat model since no assumption is made about their behavior. Second, most of the current genomic data systems assume static databases, where the datasets are published and are rarely or never updated. In practice, databases can be updated by the addition or removal of information. However, in practice, and considering that the data owner should be able to revoke the access permissions to his/her data at any time, the databases must allow not only the addition of information, but also the removal of some data without privacy breaches. This is currently an open challenge. Last, ensuring the interoperability between data from different sources requires standardization of the data formats along with the definition of standards for genomic data privacy. However, currently, the legislation for genomic data protection is not globalized. So far, each country defines the

laws related to genomic data privacy. This complicates data sharing and collaborations between geodistributed institutions. However, the research community is aware of the need for improving this field. Although some entities have been working in the field, such as Global Alliance (GA4GH), trying to establish some common guidelines and define standard methods for genomic data protection.

6. Conclusion and future work

In the context of DNA computing, some applications make use of synthesized DNA that is similar to the human DNA; thus, this chapter discussed the properties of human DNA and privacy attacks that consider it. Genomic data is used in multiple fields due to its informative nature; however, it encodes highly sensitive and personal information that is unique for each individual. Therefore, privacy is essential on the genomic data life cycle, i.e., storage, sharing, and processing. Furthermore, DNA information leakage can result in unwanted and irreversible harm. Once it is leaked, genomic privacy cannot be recovered since DNA is nonrevocable. Moreover, intra- and intergenomic data correlations, respectively, among different regions in a DNA sequence and among family members can leak additional information, including hidden information if not adequately handled. The multiple privacy attacks reported on genomic data demonstrated the main vulnerabilities of current processing algorithms which would not provide enough privacy protection to the data. This urged the development of privacy-preserving approaches. The main challenges raised in this field are mainly the protection of data privacy and the practical performance, since often privacy protection requires performance sacrifices. Privacy-preserving techniques have been used to allow multiple genomic data applications, although some applications still remain unprotected. One of the main reasons for these limitations is the performance and/or data utility sacrifice most of the privacy-preserving techniques imply in order to improve privacy protection.

Genomic data privacy research has to address several open challenges. The development of privacy-preserving systems that are able to tolerate malicious adversaries, that intent to explore vulnerabilities to gain unauthorized access to genomic data, is important since this threat model is more realistic than the often assumed honest-but-curious adversary. In addition, the improvement of current privacy-protection techniques to enable updates on the data while guaranteeing the adequate privacy protection would also

be a valuable advance in the field. More dynamic datasets would allow the production of more accurate statistics to speed up research. Finally, standardization would benefit the genomic data privacy and allow better interoperability between data from different sources, for example, from different studies that collect data in a geodistributed fashion. Currently per-country laws limit the large potential of genomic data sharing and research collaborations.

References

[1] M. Naveed, E. Ayday, E.W. Clayton, J. Fellay, C.A. Gunter, J.P. Hubaux, B.A. Malin, X. Wang, Privacy in the genomic Era, ACM Comput. Surv. 48 (1) (2015) 1–44.

[2] P.E. Verissimo, A. Bessani, E-biobanking: what have you done to my cell samples? Secur. Priv. 11 (6) (2013) 62–65.

[3] A. Bessani, J. Brandt, M. Bux, V. Cogo, L. Dimitrova, J. Dowling, A. Gholami, K. Hakimzadeh, M. Hummel, M. Ismail, E. Laure, U. Leser, J.E. Litton, R. Martinez, S. Niazi, J. Reichel, K. Zimmermann, BiobankCloud: a platform for the secure storage, sharing, and processing of large biomedical data sets, in: Proceedings of the Big-O(Q)/DMAH@VLDB 2015, 2015, pp. 86–105.

[4] M. Fernandes, J. Decouchant, F.M. Couto, P. Esteves-Veríssimo, Cloud-assisted read alignment and privacy, in: Proceedings of the 11th International Conference on Practical Applications of Computational Biology & Bioinformatics, 2017.

[5] A. Michalas, N. Paladi, C. Gehrmann, Security aspects of e-health systems migration to the cloud, in: Proceedings of the IEEE 16th International Conference on e-Health Networking, Applications and Services, 2014, pp. 212–218.

[6] B. Fabian, T. Ermakova, P. Junghanns, Collaborative and secure sharing of healthcare data in multi-clouds, Inf. Syst. 48 (2015) 132–150.

[7] M. Humbert, E. Ayday, J.P. Hubaux, A. Telenti, Addressing the concerns of the lacks family: quantification of kin genomic privacy, in: Proceedings of the ACM SIGSAC Conference on Computer & Communications Security, 2013, pp. 1141–1152.

[8] S. Gottlieb, US employer agrees to stop genetic testing, Br. Med. J. 322 (7284) (2001) 449.

[9] M. Fernandes, Reconciling data privacy with sharing in next-generation genomic workflows, (PhD thesis), University of Luxembourg 2020.

[10] E. Shapiro, T. Ran, Molecules reach consensus, Nat. Nanotechnol. 8 (2013) 703–705.

[11] R. Klitzman, P.S. Appelbaum, W. Chung, Should life insurers have access to genetic test results? JAMA 312 (18) (2014) 1855–1856.

[12] A.M.Y. Goh, E. Chiu, O. Yastrubetskaya, C. Erwin, J.K. Williams, A.R. Juhl, J.S. Paulsen, Perception, experience, and response to genetic discrimination in Huntington's disease: the Australian results of The International RESPOND-HD study, Genet. Test. Mol. Biomarkers 17 (2) (2013) 115–121.

[13] D.R. Nyholt, C.-E. Yu, P.M. Visscher, On Jim Watson's APOE status: genetic information is hard to hide, Eur. J. Hum. Genet. 17 (2009) 147–149.

[14] R. Wang, Y.F. Li, X. Wang, H. Tang, X. Zhou, Learning your identity and disease from research papers: information leaks in genome wide association study, in: Proceedings of the ACM Conference on Computer and Communications Security, 2009, pp. 534–544.

[15] J. Gitschier, Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project, Am. J. Hum. Genet. 84 (2) (2009) 251–258.

[16] E.E. Schadt, S. Woo, K. Hao, Bayesian method to predict individual SNP genotypes from gene expression data, Nat. Genet. 44 (5) (2012) 603–608.
[17] S.S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.P. Hubaux, Z. Kutalik, Quantifying genomic privacy via inference attack with high-order SNV correlations, in: Proceedings of the IEEE Security and Privacy Workshops, 2015, pp. 32–40.
[18] E. Ayday, M. Humbert, Inference attacks against kin genomic privacy, IEEE Secur. Priv. 15 (5) (2017) 29–37.
[19] P. Berrang, M. Humbert, Y. Zhang, I. Lehmann, R. Eils, M. Backes, Dissecting privacy risks in biomedical data, in: Proceedings of the IEEE European Symposium on Security and Privacy, 2018, pp. 62–76.
[20] Z. He, J. Yu, J. Li, Q. Han, G. Luo, Y. Li, Inference attacks and controls on genotypes and phenotypes for individual genomic data, in: Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, pp. 930–937.
[21] B. Malin, L. Sweeney, Determining the identifiability of DNA database entries, in: AMIA Symposium, 2000, pp. 537–541.
[22] B. Malin, L. Sweeney, How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems, J. Biomed. Inform. 37 (3) (2004) 179–192.
[23] Z. Lin, A.B. Owen, R.B. Altman, Genomic research and human subject privacy, Science 305 (5681) (2004) 183–183.
[24] M.T. Goodrich, The mastermind attack on genomic data, in: Proceedings of the 30th IEEE Symposium on Security and Privacy, 2009, pp. 204–218.
[25] L. Sweeney, Simple demographics often identify people uniquely, Health 671 (2000) 1–34.
[26] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference, Science 339 (6117) (2013) 321–324.
[27] L. Sweeney, A. Abu, J. Winn, Identifying Participants in the Personal Genome Project by Name, Data Privacy Lab, IQSS, Harvard University, 2013.
[28] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, J.P. Hubaux, De-anonymizing genomic databases using phenotypic traits, Privacy Enhanc. Technol. 2015 (2) (2015) 99–114.
[29] Y. Wang, X. Wu, X. Shi, Infringement of Individual Privacy Via Mining Differentially Private GWAS Statistics, in: Proceedings of the International Conference on Big Data Computing and Communications, 2016, pp. 355–366.
[30] C. Lippert, R. Sabatini, M.C. Maher, E.Y. Kang, S. Lee, O. Arikan, A. Harley, A. Bernal, P. Garst, V. Lavrenko, K. Yocum, T. Wong, M. Zhu, W.Y. Yang, C. Chang, T. Lu, C.W.H. Lee, B. Hicks, S. Ramakrishnan, H. Tang, C. Xie, J. Piper, S. Brewerton, Y. Turpaz, A. Telenti, R.K. Roby, F.J. Och, J.C. Venter, Identification of individuals by trait prediction using whole-genome sequencing data, Natl. Acad. Sci. 114 (38) (2017) 1–6.
[31] S. Zaaijer, A. Gordon, D. Speyer, R. Piccone, S.C. Groen, Y. Erlich, Rapid re-identification of human samples using portable DNA sequencing, eLife 6 (e27798) (2017) 1–17.
[32] Y. Erlich, T. Shor, I. Pe'er, S. Carmi, Identity inference of genomic data using long-range familial searches, Science 362 (6415) (2018) 690–694.
[33] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, S.F. Nelson, D.W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS Genet. 4 (8) (2008) 1–9.
[34] R. Braun, W. Rowe, C. Schaefer, J. Zhang, K. Buetow, Needles in the haystack: identifying individuals present in pooled genomic data, PLoS Genet. 5 (10) (2009) 1–8.

[35] K.B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D.J. Hunter, J. Paschal, T.A. Manolio, M. Tucker, R.N. Hoover, G.D. Thomas, S.J. Chanock, N. Chatterjee, A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies, Nat. Genet. 41 (11) (2009) 1253–1257.
[36] S. Sankararaman, G. Obozinski, M.I. Jordan, E. Halperin, Genomic privacy and limits of individual detection in a pool, Nat. Genet. 41 (9) (2009) 965–967.
[37] D. Clayton, On inferring presence of an individual in a mixture: a Bayesian approach, Biostatistics 11 (4) (2010) 661–673.
[38] S. Shringarpure, C. Bustamante, Privacy risks from genomic data-sharing beacons, Am. J. Hum. Genet. 97 (5) (2015) 631–646.
[39] R. Cai, Z. Hao, M. Winslett, X. Xiao, Y. Yang, Z. Zhang, S. Zhou, Deterministic identification of specific individuals from GWAS results, Bioinformatics 31 (11) (2015) 1701–1707.
[40] M. Backes, P. Berrang, M. Humbert, P. Manoharan, Membership privacy in MicroRNA-based studies, in: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 319–330.
[41] N. von Thenen, E. Ayday, A.E. Cicek, Re-identification of individuals in genomic data-sharing beacons via allele inference, Bioinformatics 35 (3) (2018) 365–371.
[42] A. Kong, G. Masson, M.L. Frigge, A. Gylfason, P. Zusmanovich, G. Thorleifsson, P.I. Olason, A. Ingason, S. Steinberg, T. Rafnar, P. Sulem, M. Mouy, F. Jonsson, U. Thorsteinsdottir, D.F. Gudbjartsson, H. Stefansson, K. Stefansson, Detection of sharing by descent, long-range phasing and haplotype imputation, Nat. Genet. 40 (9) (2008) 1068–1075.
[43] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, Privacy in pharmacogenetics: an end-to-end case study of personalized Warfarin dosing, in: Proceedings of the USENIX Security Symposium, 2014, pp. 17–32.
[44] I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, E. Ayday, An inference attack on genomic data using kinship, complex correlations, and phenotype information, IEEE/ACM Trans. Comput. Biol. Bioinform. 15 (4) (2018) 1333–1343.
[45] M. Akgün, An active genomic data recovery attack, Balkan J. Elect. Comput. Eng. 7 (2019) 417–423.
[46] M.D. Edge, G. Coop, Attacks on genetic privacy via uploads to genealogical databases, eLife 9 (2020) e51810.
[47] P. Ney, L. Ceze, T. Kohno, Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference, in: Proceedings of the Network and Distributed System Security Symposium, 2020.
[48] L. Sweeney, k-anonymity: a model for protecting privacy, Int. J. Uncertainty Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.
[49] K. El Emam, F.K. Dankar, Protecting privacy using k-anonymity, J. Am. Med. Inform. Assoc. 15 (5) (2008) 627–637.
[50] E. Jonker, T. Roffey, R. Vaillancourt, J. Bottomley, E. Cogo, F.K. Dankar, K. El Emam, S. Chowdhury, D. Amyot, R. Issa, J.P. Corriveau, M. Walker, A globally optimal k-anonymity method for the de-identification of health data, J. Am. Med. Inform. Assoc. 16 (5) (2009) 670–682.
[51] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, ACM Trans. Knowl. Discov. Data 1 (1) (2007) 3–54.
[52] N. Li, T. Li, S. Venkatasubramanian, t-Closeness: privacy beyond k-anonymity and l-diversity, in: Proceedings of the IEEE International Conference on Data Engineering, 2007, pp. 106–115.
[53] B.A. Malin, Protecting DNA sequence anonymity with generalization lattices, Methods Inf. Med. 44 (2005) 687–692.

[54] Z. Lin, M. Hewett, R.B. Altman, Using binning to maintain confidentiality of medical data, in: Proceedings of the AMIA Symposium, 2002, pp. 454–458.
[55] B.A. Malin, An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future, J. Am. Med. Inform. Assoc. 12 (1) (2005) 28–34.
[56] E.C. Hayden, Privacy protections: the genome hacker. Yaniv Erlich shows how research participants can be identified from 'anonymous' DNA, Nature 497 (7448) (2013) 172–174.
[57] G. Li, Y. Wang, X. Su, Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices, Comput. Methods Programs Biomed. 108 (1) (2012) 1–9.
[58] V.V. Cogo, A. Bessani, F.M. Couto, P. Verissimo, A high-throughput method to detect privacy-sensitive human genomic data, in: Proceedings of the ACM Workshop on Privacy in the Electronic Society, 2015, pp. 101–110.
[59] J. Decouchant, M. Fernandes, M. Völp, F.M. Couto, P. Esteves-Veríssimo, Accurate filtering of privacy-sensitive information in raw genomic data, J. Biomed. Inform. 82 (2018) 1–12.
[60] M. Fernandes, J. Decouchant, M. Völp, F.M. Couto, P. Esteves-Verissimo, DNA-SeAl: sensitivity levels to optimize the performance of privacy-preserving DNA alignment, IEEE J. Biomed. Health Inform. 24 (3) (2020) 907–915.
[61] E. Vayena, U. Gasser, Between openness and privacy in genomics, PLoS Med. 13 (1) (2016) 1–7.
[62] Y. Erlich, A. Narayanan, Routes for breaching and protecting genetic privacy, Nat. Rev. Genet. 15 (2014) 409–421.
[63] J. Baron, K. El Defrawy, K. Minkovich, R. Ostrovsky, E. Tressler, 5pm: secure pattern matching, in: Proceedings of the International Conference on Security and Cryptography for Networks, 2012, pp. 222–240.
[64] M.J. Atallah, F. Kerschbaum, W. Du, Secure and private sequence comparisons, in: Proceedings of the ACM Workshop on Privacy in the Electronic Society, 2003, pp. 39–44.
[65] M. Kantarcioglu, W. Jiang, Y. Liu, B. Malin, A cryptographic approach to securely share and query genomic sequences, IEEE Trans. Inf. Technol. Biomed. 12 (5) (2008) 606–617.
[66] M. Namazi, J.R. Troncoso-Pastoriza, F. Perez-Gonzalez, Dynamic privacy-preserving genomic susceptibility testing, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2016, pp. 45–50.
[67] G.S. Çetin, H. Chen, K. Laine, K. Lauter, P. Rindal, Y. Xia, Private queries on encrypted genomic data, BMC Med. Genomics 10 (2) (2017) 45.
[68] D. He, N.A. Furlotte, F. Hormozdiari, J.W.J. Joo, A. Wadia, R. Ostrovsky, A. Sahai, E. Eskin, Identifying genetic relatives without compromising privacy, Genome Res. 24 (4) (2014) 664–672.
[69] S. Namasudra, D. Devi, S. Choudhary, R. Patan, S. Kallam, Security, privacy, trust, and anonymity, in: S. Namasudra, G.C. Deka (Eds.), Advances of DNA Computing in Cryptography, Chapman & Hall/CRC, 2018, pp. 138–150.
[70] S. Namasudra, G.C. Deka, R. Bali, Applications and future trends of DNA computing, in: S. Namasudra, G.C. Deka (Eds.), Advances of DNA Computing in Cryptography, Chapman & Hall/CRC, 2018, pp. 181–192.
[71] Y. Huang, Secure multi-party computation, in: Responsible Genomic Data Sharing, X. Jiang and H. Tang, Eds., Academic Press, 2020, pp. 123–134.
[72] M.M. Al Aziz, M.Z. Hasan, N. Mohammed, D. Alhadidi, Secure and efficient multi-party computation on genomic data, in: Proceedings of the International Database Engineering & Applications Symposium, 2016, pp. 278–283.

[73] H. Cho, D.J. Wu, B. Berger, Secure genome-wide association analysis using multiparty computation, Nat. Biotechnol. 36 (6) (2018) 547–551.
[74] D. Deuber, C. Egger, K. Fech, G. Malavolta, D. Schröder, S.A.K. Thyagarajan, F. Battke, C. Durand, My genome belongs to me: controlling third party computation on genomic data, Proc. Priv. Enhanc. Technol. 2019 (1) (2019) 108–132.
[75] A. Mittos, B. Malin, E.D. Cristofaro, Systematizing genome privacy research: a privacy-enhancing technologies perspective, Priv. Enhanc. Technol. 2019 (1) (2019) 87–107.
[76] K. Learned, A. Durbin, R. Currie, E.T. Kephart, H.C. Beale, L.M. Sanders, J. Pfeil, T.C. Goldstein, S.R. Salama, D. Haussier, O.M. Vaske, I.M. Bjork, Barriers to accessing public cancer genomic data, Sci. Data 6 (98) (2019) 907–915.
[77] Y. Erlich, J.B. Williams, D. Glazer, K. Yocum, N. Farahany, M. Olson, A. Narayanan, L.D. Stein, J.A. Witkowski, R.C. Kain, Redefining genomic privacy: trust and empowerment, PLoS Biol. 12 (11) (2014) 1–5.
[78] C.C. Agbo, Q.H. Mahmoud, J.M. Eklund, Blockchain technology in healthcare: a systematic review, Healthcare 7 (2) (2019) 56.
[79] M. Hölbl, M. Kompara, A. Kamisalic, L. Nemec Zlatolas, A systematic review of the use of blockchain in healthcare, Symmetry 10 (470) (2018).
[80] S. Namasudra, Fast and secure data accessing by using DNA computing for the cloud environment, IEEE Trans. Serv. Comput. 15 (4) (2022) 2289–2300.
[81] F. Rocha, M. Correia, Lucy in the sky without diamonds: stealing confidential data in the cloud, in: Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks Workshops, 2011, pp. 129–134.
[82] E.S. Dove, Y. Joly, A.M. Tasse, Public Population Project in Genomics and Society (P3G) International Steering Committee and International Cancer Genome Consortium (ICGC) Ethics and Policy Committee, B.M. Knoppers, Genomic cloud computing: legal and ethical points to consider, Eur. J. Human Genet. 23 (2015) 1271–1278.
[83] X. Zhou, B. Peng, Y.F. Li, Y. Chen, H. Tang, X. Wang, To release or not to release: evaluating information leaks in aggregate human-genome data, in: Proceedings of the European Symposium on Research in Computer Security, 2011, pp. 607–627.
[84] M. Blanton, M.J. Atallah, K.B. Frikken, Q. Malluhi, Secure and efficient outsourcing of sequence comparisons, in: European Symposium on Research in Computer Security, 2012, pp. 505–522.
[85] Y. Chen, B. Peng, X. Wang, H. Tang, Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds, in: Proceedings of the Network & Distributed System Security Symposium, 2012.
[86] V. Popic, S. Batzoglou, A hybrid cloud read aligner based on MinHash and kmer voting that preserves privacy, Nat. Commun. 8 (15311) (2017) 1–7.
[87] A. Bessani, M. Correia, B. Quaresma, F. André, P. Sousa, DepSky: dependable and secure storage in a cloud-of-clouds, ACM Trans. Storage 9 (4) (2013) 1–33.
[88] R. Mendes, T. Oliveira, V.V. Cogo, N.F. Neves, A.N. Bessani, CHARON: a secure cloud-of-clouds system for storing and sharing big data, IEEE Trans. Cloud Comput. 9 (4) (2021) 1349–1361.
[89] J.L. Raisaro, C. Troncoso, M. Humbert, Z. Kutalik, A. Telenti, J.P. Hubaux, GenoShare: supporting privacy-informed decisions for sharing exact genomic data, EPFL Infoscience (2017) 1–19.
[90] V. Cogo, A. Bessani, Enabling the efficient, dependable cloud-based storage of human genomes, in: Proceedings of the International Symposium on Reliable Distributed Systems Workshops, 2019.
[91] M. Schwarz, S. Weiser, D. Gruss, C. Maurice, S. Mangard, Malware guard extension: using SGX to conceal cache attacks, in: In Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, 2017, pp. 3–24.

[92] J. Götzfried, M. Eckert, S. Schinzel, T. Müller, Cache attacks on Intel SGX, in: Proceedings of the European Workshop on Systems Security, 2017, pp. 1–6.
[93] F. Chen, C. Wang, W. Dai, X. Jiang, N. Mohammed, M.M. Al Aziz, M.N. Sadat, C. Sahinalp, K. Lauter, S. Wang, PRESAGE: PRivacy-preserving gEnetic testing via SoftwAre Guard Extension, BMC Med. Genomics 10 (48) (2017) 77–85.
[94] F. Chen, S. Wang, X. Jiang, S. Ding, Y. Lu, J. Kim, S.C. Sahinalp, C. Shimizu, J.C. Burns, V.J. Wright, E. Png, M.L. Hibberd, D.D. Lloyd, H. Yang, A. Telenti, C.S. Bloss, D. Fox, K. Lauter, L. Ohno-Machado, PRINCESS: Privacy-protecting rare disease International Network Collaboration via Encryption through Software guard extensionS, Bioinformatics 33 (6) (2017) 871–878.
[95] C. Lambert, M. Fernandes, J. Decouchant, P. Esteves-Veríssimo, MaskAl: Privacy Preserving Masked Reads Alignment using Intel SGX, in: Symposium on Reliable Distributed Systems (SRDS), 2018.
[96] M. Völp, J. Decouchant, C. Lambert, M. Fernandes, P. Esteves-Verissimo, Enclave-based privacy-preserving alignment of raw genomic information: information leakage and countermeasures, in: Proceedings of the 2nd Workshop on System Software for Trusted Execution, 2017, pp. 1–6.
[97] B. Zubairu, Security risks of biomedical data processing in cloud computing environment, in: Cloud Security: Concepts, Methodologies, Tools, and Applications, Information Resources Management Association, 2019, pp. 1748–1768.
[98] T. Pascoal, J. Decouchant, A. Boutet, P. Esteves-Verissimo, DyPS: Dynamic, Private and Secure GWAS, in: Proceedings on Privacy Enhancing Technologies, Sciendo, 2021.
[99] K. Ayoz, E. Ayday, A.E. Cicek, Genome reconstruction attacks against genomic data-sharing beacons, arXiv preprint:2001.08852 (2020).

About the authors

Maria Fernandes is currently a postdoctoral researcher at University of Oxford. She received a master in Bioinformatics and Computational Biology from University of Lisbon, and a PhD degree in Computer Science from the University of Luxembourg. Her research interests focus on biomedical data storage and analysis, data privacy and human genetics. During her PhD, she worked on designing privacy-preserving approaches for genomic data processing.

Jérémie Decouchant is an assistant professor at Delft University of Technology (NL). Previously, he has been a research scientist at SnT, University of Luxembourg. He received an engineering degree (MSc) from Ensimag, Grenoble, and a PhD degree in computer science from the University of Grenoble-Alpes, France. His research interests revolve around resilient distributed systems and algorithms. In particular, he has been designing privacy-preserving processing workflows for genomic data.

Francisco M. Couto is currently an associate professor with habilitation at Universidade de Lisboa (Faculty of Sciences) and a researcher at LASIGE. He graduated (2000) and has a master (2001) in Informatics and Computer Engineering from the IST. He concluded his doctorate (2006) in Informatics, specialization Bioinformatics, from the Universidade de Lisboa. He was an invited researcher at EBI, AFMB-CNRS, BioAlma during his doctoral studies. His main research contributions cover several key aspects of bioinformatics and knowledge management, namely in proposing and developing: various text mining solutions that explore the semantics encoded in ontologies; semantic similarity measures and tools using biomedical ontologies; and ontology and linked data matching systems. Until August 2022, he published 2 books; was co-author of 10 chapters, 62 journal papers (47 Q1 Scimago), and 32 conference papers (10 core A and A*); and was the supervisor of 10 PhD theses and of 51 master theses. He received the Young Engineer Innovation Prize 2004 from the Portuguese Engineers Guild, and an honorable mention in 2017 and the prize in 2018 of the ULisboa/Caixa Geral de Depósitos (CGD) Scientific Prizes.