Collaborative analysis of genomic data: vision and challenges

Sara Jafarbeiki*†, Raj Gaire†, Amin Sakzad*, Shabnam Kasra Kermanshahi‡, Ron Steinfeld*

*Faculty of Information Technology, Monash University, Australia
†CSIRO Data61, Australia
‡School of Computing Technologies, RMIT University, Australia
{sara.jafarbeiki, amin.sakzad, ron.steinfeld}@monash.edu
{sara.jafarbeiki, raj.gaire}@data61.csiro.au
shabnam.kasra.kermanshahi@rmit.edu.au

Abstract—The cost of DNA sequencing has resulted in a surge of genetic data being utilised to improve scientific research, clinical procedures, and healthcare delivery in recent years. Since the human genome can uniquely identify an individual, this characteristic also raises security and privacy concerns. In order to balance the risks and benefits, governance mechanisms including regulatory and ethical controls have been established, which are prone to human errors and create hindrance for collaboration. Over the past decade, technological methods are also catching up that can support critical discoveries responsibly. In this paper, we explore regulations and ethical guidelines and propose our visions of secure/private genomic data storage/processing/sharing platforms. Then, we present some available techniques and a conceptual system model that can support our visions. Finally, we highlight the open issues that need further investigation.

Index Terms—Genomic data, privacy and security, collaborative computing, regulations and ethics

I. INTRODUCTION

Thanks to decreasing cost of sequencing genomes, the acquisition of genomic data are increasing. These data help researchers and clinicians in biomedical research/analysis and clinical procedures. This advancement is improving existing healthcare practices such as giving care after an illness and helping to anticipate, prescribe, and prevent illnesses from occurring in the first place. In other words, these analyses enable the delivery of evidence-based healthcare.

Genomic data have specific characteristics that make them sensitive to concerns of privacy and security. This concern is mainly due to special features of genomic data, which include (i) a link between characteristics and specific diseases, (ii) the capability to identify individuals, and (iii) the disclosure of family relationships. For example, if a person has mutations in BRCA genes, she is more likely to develop breast cancer [1]. Due to the impact of data leakage on personal life [2], [3], including job loss due to medical history, no employment, and excessive insurance premiums, care must be taken at all phases of the genomic data generation, storage, and use. Furthermore, the owner's consent for the use and reuse of the data becomes important, and hence the governance of data becomes critical.

Current governance based mechanisms (see Section II) mostly include policies and legal instruments that rely on human based processes. Although these processes work in most situations, they are prone to human errors, slowness due to the necessary bureaucracy and hindrance for collaboration. Technologies can help assure the adequate processes are followed with less errors and at the speed that would allow collaborations among researcher to foster towards new discoveries.

There are surveys that cover various techniques to address the privacy and security requirements of genomic data, such as [4]–[7]. However, there is no work to inclusively, (i) identify the current state of genomic data privacy and security solutions, (ii) see the future visions, and (iii) connect them with the enablers (techniques). Therefore, in this paper, we explore a vision for the future of genome analysis, analyse how technologies can contribute to achieving those visions, and the open issues that still need further research and development. While governance has a role in genomic data privacy/security assurance, our vision is not to rely on policies (or non-technical aspects) alone. By putting more effort into techniques, such as cryptography, novel approaches can be developed that can assist humans and guarantee the required security and privacy requirements. Before getting into those details, let us first look into some background on genomic data to establish a context on the subject matter.

A. Genomic data and its characteristics

The genome shows an organism's overall genetic information, encoded in deoxyribonucleic acid molecules know as DNA. DNA consists of two long complementary polymer chains of four units called nucleotides. These nucleotides are represented by the acronyms A, C, G, T (Adenine, Cytosine, Guanine and Thymine). It is a chain-like molecule of variable length made of these letters. It includes details on a person's health and behavior. It does not change much over time. Although most of the human genome look alike, small changes across the large genome make it unique for each individual. That means any two people's DNA can be easily distinguished from one another. Moreover, an individual's DNA carries information about her biological relatives and leaks information beyond the individual from whom it was acquired [8]. It can be used to recognise disease predisposition and physical traits that make DNA a large collection of sensitive information. Such

knowledge about a person can have significant implications [9], [10] and is vulnerable to privacy risks and attacks [4].

B. Genomic data collection process

The process of genomic data collection is shown in Fig. 1. As illustrated, patients and also healthy people (as data owners) provide their samples (usually blood samples) to a medical institution (or hospital or a gene trustee). They also provide consent for different uses of their samples, including the usage of information available in their samples. The medical institutions collect, sequence (i.e., determine characters of DNA) and process (i.e., clean up and prepare for further analysis) the genomic data. Then, they manage the information (genomic data, diseases, metadata such as gender, ethnicity, age and consent) and store the data in a biobank. For meaningful analysis, researchers need to test their hypotheses on a large number of samples. Therefore, these data are shared in public databases, biobanks and repositories (e.g., UK biobank [11] and the 1000 Genomes Project [12]).

C. Uses of genomic data

There are several usage of genomic data [4]. Here, we focus on two main usages: personalised health and disease research.

Personalised Health: It has long been known that changes in a person's genetic sequence can have an impact on their health. Changes in a particular gene can have a negative impact on a person's health instantly or in the future [13]. While some of these diseases are treatable with dietary modifications or pharmaceutical therapies, others are not, and there is no proven intervention to help improve a person's health. Nonetheless, some people want to know their genetic status to prepare for the expected future and contribute to medical research. Genetic testing has applications in areas such as (i) paternity testing, to decide if two individuals are parents and offspring, (ii) diagnostic testing, to determine if an individual is impaired by a specific disease and (iii) pharmacogenetic testing, in personalised medicine to customise patient care and to develop successful drugs and therapies [14].

Disease Research: Despite the fact that the genome has been connected to a large number of diseases and treatment responses, new connections are being identified regularly. The technology for doing basic research is rapidly evolving [15]. Today's amount of genetic data being gathered, stored, and processed is extraordinary, hastening the identification of new mutations. Meaningful discoveries require a large number of relevant samples to ensure the statistical significance of the results. Since gathering such a large number of samples in a single institution may not be possible, the data need to be shared among researchers in different institutes. As more collaboration is needed on genomic data for research, so too will the privacy and security assurance of genomic data [16].

Due to the sensitivities of these data, several governance mechanisms, including legal and ethical instruments, have been established to ensure the data is shared and used in a responsible way.

II. GOVERNANCE OF GENOMIC DATA

The genome data has special features and cannot be treated like other data. Hence, it is important to consider protective measures before sharing the genomic data, to provide privacy of the individuals who contribute the data and reduce harm from the data. Adhering to the law and ethics in one of such important considerations.

A. Law and Regulations

The law is a set of rules that a country or society acknowledges as governing its people's activities and that it may enforce through the application of penalties [17]. It is governed by a government. Ethics, on the other hand, are moral rules that regulate a person's behavior or active conduct [18]. Several nations and organizations have taken major steps in enacting laws and regulations governing health data privacy, security, and trust.

The Privacy Act 1988 [19] governs the privacy and security framework for personal information in Australia. Most Commonwealth government agencies (including the tax office and the Department of Human Services), all private sector organizations with an annual turnover of more than three million dollars, and a few other organizations that meet specific criteria, such as health service providers, are covered by the Privacy Act. The Privacy Act 1988 ('Privacy Act') defines 'health information' as including 'genetic information about an individual in a form that is, or could be, predictive of the health of the individual or a genetic relative of the individual'.

GINA (Genetic Information Nondiscrimination Act of 2008) [20] and, more subsequently, HIPAA (Health Insurance Porta-
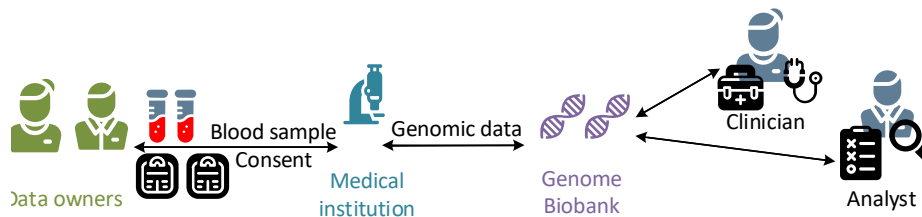

Fig. 1. Genomic data from generation to use in medical analysis/care

bility and Accountability Act of 1996) [21] are laws that deal with genetic information. GINA is an anti-discrimination law. It prohibits group health and Medicare supplementary plans from using genetic information to discriminate against individuals when it comes to insurance, but not life, disability, or long-term care policies. It also prohibits businesses from utilising genetic data in hiring, firing, promotions, or work assignments. Similarly, HIPAA provides guidelines about how to handle sensitive information through de-identification. HIPAA laws were updated in 2013 by the HIPAA Omnibus Rule, which included genetic information in the scope of Protected Health Information (PHI).

The General Data Protection Regulation (GDPR) [22], a new framework for European data privacy legislation, was authorised and enacted by the European Commission in 2016. The GDPR protects individuals' privacy rights and aims to harmonize data protection rules across Europe, regardless of where data is handled. The GDPR and genomic data report [23] examines how the GDPR affects genomic healthcare and research, highlights areas that require attention, and gives recommendations to the genomics community.

Some of the overall requirements due to these regulations are discussed in Table I.

B. Ethics

In Australia, the laws governing genetic data exchange are a mix of common law, legislation, ethical principles, and codes of practice. The primary regulatory lever is the National Health and Medical Research Council's (NHMRC) National Statement for Ethical Conduct in Human Research (National Statement) [25], which applies to genetic data that satisfies the criteria of personal information.

The Indian Council of Medical Research (ICMR) [26] has provided comprehensive guidelines for Indian researchers working in the genomic area, and a separate chapter has been included in the 2006 issue of "Ethical guidelines for biomedical research on human subjects." The prospect of re-identification of users through connecting, combining, data-mining, and reusing large databases raises ethical concerns.

There are other ethical studies such as [27] which is performed by an Italian company named Evodevo srl, and is supported by the European Economic and Social Committee. It explores the ethical requirements of big data, which also includes health and genomic data. Another ethics guideline can be found in [28], which is based on the American Medical Association code of medical ethics.

In various ethical guidelines such as the above-mentioned ones, the main points are (i) privacy and confidentiality, and (ii) consent. Apart from these two, ethics approval in research projects, data sharing management, and risk analysis and management are also discussed.

Legal and ethical aspects in genomic data security and privacy provide some ways of data protection. Different health data security/privacy issues have been solved using policies/regulations. For example, *consent* is taken from data owners for different uses of their genomic data using papers with all the information included to be signed. This does not resolve different requirements of consent revocation/update thoroughly. Moreover, when using genomic data for analysis to *generate knowledge*, the dataset is being downloaded by authorised researchers/analysts. The problem is that the consent revocation is hard in this case and after downloading data, it would be nontransparent. Furthermore, for more accurate analysis, when data is stored in different organisations, *interconnection and integration* becomes a key. It is done in an institutional way, where a consortium is created and people signed up to that consortium are able to share data and access it. In addition to all the above, *transparency and accountability* are essential when working with sensitive genomic data. The data is distributed to analysts based on policies/regulations and penalties are enforced to someone who does not obey them.

TABLE I
OVERALL REQUIREMENTS DUE TO LAW/REGULATIONS

Overall requirements	Regulations		
	Privacy Act [19]	*GDPR* [22]	*HIPAA* [21]
Confidentiality, integrity and availability of health data	Security and integrity are discussed.	Data security and end-to-end encryption are considered.	More Health data must be kept confidential, integral, and available.
Transparency and fairness	Open and transparent handling of personal information are required, as well as anonymity and pseudonymity.	Lawfulness, transparency and fairness are considered when handling personal health data.	Policies, methods, and technology that directly influence individuals and their health information should be open and transparent.[a]
Consent	Consent is required for disclosing personal health data and it can be only used for that purpose consent has been given for. Consent can be withdrawn by individuals.	The aim of data use must be clearly stated, unambiguous, and legal. The health data has to be used for reasons for which consent (can be withdrawn) was collected.	Consent is required for secondary use of health data, and in a written form. Consent can be revoked by individuals in writing.
Access control	Individual access can be provided on request.	Protect health data against unauthorised or unlawful processing and accidental loss or damage.	Access has to be authorised, e.g., considering access control with unique user identification.
Breach notification	Affected individuals and the Australian Information Commissioner have to be notified about a breach.	Individuals have right to be notified if their data is breached within 72 hours.	Provide notification following a breach of an individual's personal health data.

[a]The U.S. Department of Health and Human Services (HHS) reformed the requirements [24].

III. PROBLEMS

By relying only on regulations and ethics to provide guidelines, some problems arise that are explained in this section.

A. Human error

When human is involved in the genomic data handling and process, human error can cause problems. Example scenarios of human errors may include (i) Sending sensitive information to the wrong recipients; (ii) Sharing password/account information; and (iii) End user and IT staff failure to follow the rules and procedures. Therefore, sensitive data is exposed as a result of data breaches [29], [30]. The healthcare business is one of the worst affected in comparison to other data industries [31]. According to several practitioners, the overall number of people affected by healthcare data breaches was 249.09 million from 2005 to 2019 [32]. Human error was also identified as a top source of Australian data breaches in 2020, where it accounted for 38% of the breaches [33] and health industries was one of the most affected industries. If a breach occurs, genetic data may be used to identify an individual and forecast their physical traits [9], [10], and familial matching can be performed by matching individuals to distant relatives [34]. Furthermore, individual and family health information can be revealed using genetic data [35].

B. Trust-based mechanisms

Upon approval, an analyst can download the requested data, and it will be her responsibility if misused. This process is being handled based on trusting the authorised users not to distribute the data or misuse it. Otherwise, penalties might be applied. However, apart from the trust requirement that can be an issue, human error may be added when data is downloaded and used. This situation leads to the data breaches mentioned in the previous subsection.

C. Hindrance to collaboration

Human involvement can make a process slow. Relying on the human to assess the requests and authorise access needs governance processes and paperwork that needs time to complete. As such, data requests may take a long time to process. This bureaucracy lowers the speed of analysis. If the data analysis process can be automated and all the checks can be done using computational techniques, not only will the process be quicker, but also the human error in authorising and assessing will be eliminated.

D. Consent management

An informed consent is required in most conditions when a medical procedure is performed. For using the data in analysis and study, to obtain consent, the subject must be informed of his or her rights, the intent of the study, the procedures to be performed, the possible risks and benefits of participation, the anticipated length of the research, the degree to which personal identity and demographic data are confidential, so that subjects' participation in the study is entirely informed and voluntary [36]. However, one-time consent is usually not sufficient and individuals should be able to revoke their consent regarding the access to their data.

In general, consent may be approached in two ways: static consent and dynamic consent. Static consent requires that an agreement, generally a paper document, is signed by the participants at the moment of data collection for all future data uses. This approach does not consider the difficulties that arise as the environment and requirements change over time, such as utilising the data for a different health initiative than the one for which the consent was given. Furthermore, written consent forms are hard to manage, especially when it comes to their revocation. In addition, when the data is shared with analysts and a revocation request comes in, managing it would be a problem. Automating the consent (that goes with the data) and managing it efficiently is still a problem.

In light of these existing issues, it is critical to investigate the requirements/techniques for genomic data privacy and security. Considering the existing problems, we discuss our future visions in the next section. Some of our visions are probably near future visions, and some need more investigations and might take time to be established.

IV. FUTURE VISIONS

This section presents our visions with respect to genomic data privacy and security. The future visions and their current states that have some issues (discussed at the end of section II) are represented in Fig. 2.
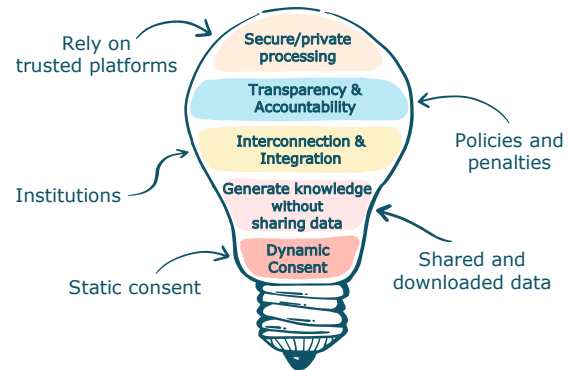

Fig. 2. Future visions

A. Secure and private processing/computing of genomic data

Computation on genomic data is an important factor in the data analysis to help researchers with statistical analysis, genome-wide association studies (GWAS), etc. There are platforms such as Google cloud confidential computing and Azure confidential computing that make it convenient, but this way, there is a need for a trusted platform. This vision can also be combined with the fourth vision, which is *generate knowledge without sharing data*. Therefore, our vision is that while the data is not shared with requesters, the computations are done in a secure/private way on the data. Different techniques have been used to provide this in the literature but they either rely

on a trusted platform (e.g., Google cloud) or their protocols (e.g., Homomorphic encryption-discussed in section V) are not ready as a product to be used. In this regard, we present our vision as there has to be computations/analysis done on the stored data that is not shared with anyone and in a secure/private way.

B. Transparency and accountability

Accountability requires transparency. To build trust, governance of genomic data must be based on transparency and accountability. It is required to maintain an accounting of all disclosures, even including those related to treatment and healthcare operations. It is difficult to determine if data users are accountable for the data entrusted to them if they do not make efforts to verify that usage limits are observed. Clear data exchange is required to guarantee stakeholder responsibility. Our vision is to provide technical ways that can support transparency and accountability without just relying on the regulations/guidelines and penalties for those who do not obey the regulations/guidelines and, e.g., share the data with others or misuse it.

C. Interconnection and integration

There are different institutions/medical centers that might have an individual's medical records/genomic information. If all the data is not in one place there would be some challenges in both analysing the data, and accessing it for medical purposes. Regarding genome analysis, if an analyst is analysing the data, they should be able to connect them and use them as big data for better output (in other words, integrate data). When it comes to the interconnection of the medical records and genomic data, the challenge goes to the interconnection of data (genomic, medical and health) that are dispersed. There are relations between genomic data and medical records and health conditions, that need to be taken into consideration while analysing or caring. Therefore, integration is much needed when the genomic data is stored in different places/hospitals/institutions. More importantly, interconnection is needed when the genomic data is stored in one institution and medical data is stored in another institution or hospital. Our vision is to integrate genomic data in a dispersed scenario, where genomic data is stored in different institutions and interconnect the genomic data and medical records/health conditions stored in different places but all in a secure/private way.

D. Generate knowledge without sharing data

Knowledge can be a specific information about a patient, the result of statistical analysis on genomic data, trained model on genomic dataset, or result of other analysis/queries on genomic data. Genomic data needs to be analysed with other information such as diseases, specifications, e.g., age. Therefore genomic data alongside other information about individuals are gathered and stored. There are two main ways of storing genomic data, centralised and decentralised storage. In centralised storage, all the data from different hospitals or institutions are collected and stored in one data server. In decentralised storage, the data are stored in different data servers. It means that there is distributed storage of genomic data, e.g., hospitals. The centralised storage has a risk of failure and attack on all the data we have. However, in decentralised storage, these risks are localised, but computation and processing of data become more complicated. With accepting the risks in both cases and providing the privacy and security requirements of both types of storage, our vision is to generate all the knowledge needed without sharing the data with analysts. It means that the request comes to the data, the result/knowledge will be generated and sent back to the requester without letting them downloading all the data.

E. Dynamic consent

Owing to the extremely diverse nature of current healthcare settings, one-time consent management approaches are insufficient to address the specific privacy issues and needs of the individual patient [37]. Dynamic consent is an informed and individualised consent in which the data subject and data custodian have two-way communication and the subject may update and grant various types of consent. In addition, the subject may use the interface to oversee their health data usage over time and revoke consent. Furthermore, the consent is transmitted with the related data when it is shared with third parties, and the participant has access to the research findings. Dynamic consent, on the other hand, has challenges, such as higher implementation costs, consent revocation, and data deletion assurance, as well as the requirement that patients have adequate digital expertise and time. Some works have already been done on consent to model the consent and manage it such as [38] that models the consent in a tool, have a repository for storing it, and a data access management feature to enforce consent. Dynamic consent management has not been implemented and used widely in hospitals/institutions/gene trustees. It is our vision that dynamic consent will be designed and used widely (while providing data security and privacy) to ensure data owners have full control over their data.

V. Techniques to support visions

In this section, we introduce some important data security and privacy-preserving techniques that are able to support our visions. A taxonomy of future visions, existing techniques to support visions, and open issues for genomic data privacy and security has been shown in Fig. 3. The relation between techniques and visions has been demonstrated by using numbers in the taxonomy. It is worth mentioning that there may be other techniques that can assist with reaching future visions, and also these techniques can be combined and used together in some cases to provide the requirements.

A. Privacy Preserving solutions

1) Differential privacy: Differential Privacy (DP) protects the privacy of output data or results from computations or processes by limiting the amount of leakage of individual input data to the authorized (usually minimal) amount. The Laplace
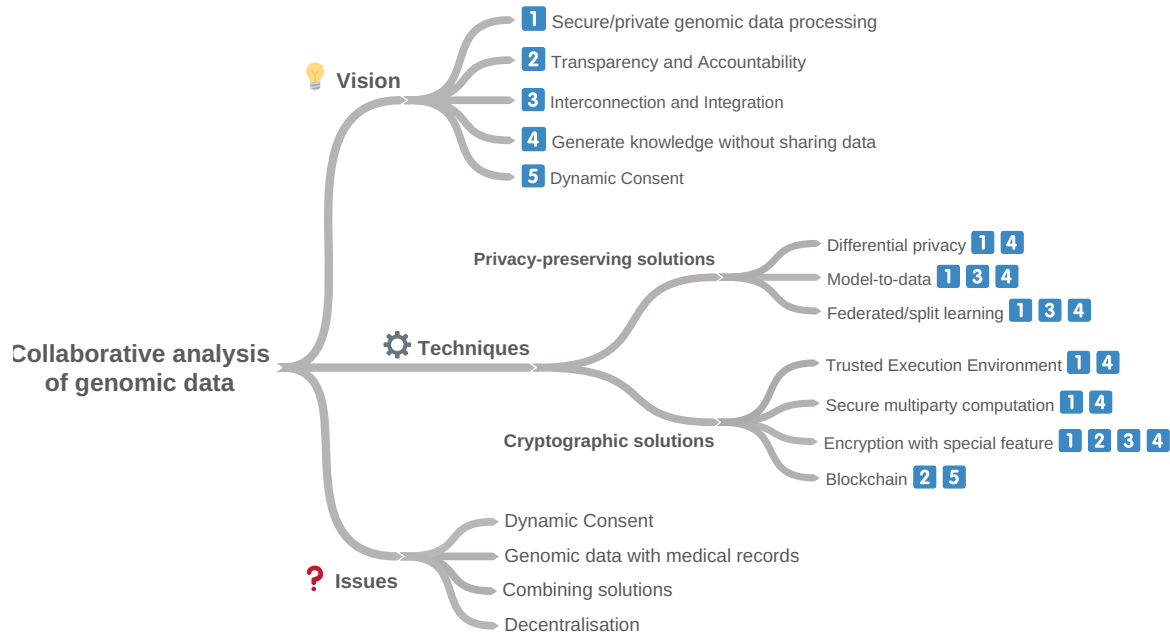
Fig. 3. A taxonomy of future visions, existing techniques to support visions, and open issues for genomic data privacy and security

mechanism [39] and the exponential mechanism [40] are the most common methods for realising differentially private algorithms, in which a random noise generated from a Laplace distribution and a scaled symmetric exponential distribution are added to the output data to achieve DP, respectively. Local and global differential privacy are the two forms of DP. Each dispersed participant adds noise to their input data before collection or computation in local DP, whereas noise is added to the final output after computation in global DP. DP is used in conjunction with an encryption method. In a dispersed clinical setting, a combination of encryption and DP is utilised to ensure the privacy of genomics data [41].

2) Model-to-data approach: Machine learning (ML) approaches are the important part of health data computation. Based on data accessibility, there are two sorts of computing techniques for the stored health data, data-to-modeler (DTM) and model-to-data (MTD) [42]. A data modeler has direct access to the data for model building (training and validation) and hypothesis testing in the Data-to-Modeler (DTM) method. Because the data modeler must be trusted, and data, which is sensitive in the healthcare sector, is directly accessible by the modeler, this technique does not follow the privacy rules since the first phases of relevant design and creation (privacy-by-design approach). The data is not directly accessible in Model-to-Data (MTD); thus a modeler must submit their codes and models (derived from their initial data) to the data provider. To put it another way, the model moves to the data. The models are then trained and verified on the unseen actual data at the data contributors' server. The modeler is notified of the revised model or outcome. MTD takes a privacy-by-

design approach. It allows for computation without access to the personal information [43].

3) Federated and Split learning: Federated learning (FL) and split learning (SL) are distributed deep learning algorithms that do not necessitate the exchange of raw data from multiple locations. FL shifts computation to edge devices. It protects the privacy of data at all participating edge devices. The training data are normally provided in one central server (or datacenter) in the classic ML technique, and the training takes place on that server, which is generally a trusted platform that can see data while computing. The FL method, on the other hand, protects the privacy of data stored at edge devices by undertaking collaborative learning that never needs data from dispersed edge devices to be acquired outside of those devices [44]. SL is particularly beneficial when data sharing across data holders/sources is impossible owing to resource constraints, as well as privacy and legal concerns. To begin, each client trains a neural network up to a certain layer, called the cut layer, and then sends the activation data from that layer to the server. The server then trains the remainder of the network layers and completes the forward propagation after receiving the activation from the client. Second, the server computes the loss using the true and predicted labels, then backpropagates the loss to calculate the correction in the form of gradients until the cut layer is reached. After that, the cut layer's gradient is sent back to the client. The client continues the network's backpropagation on its side after receiving the gradients. The forward and backward propagations will continue in this manner until the network has been trained [45].

Federated learning has been used in healthcare applications such as [46], that employs FL on brain tumor segmentation data. It has also been used to predict hospitalisation of patients who have heart diseases [47]. It has also been discussed in [48] as an expansion of their project, which is automating machine learning workflows for genomics, to enable learning across different genomic data storages.

B. Cryptographic solutions

1) Trusted execution environment: The Trusted Execution Environment (TEE) protects important computations by isolating them from other processes like as operating systems, BIOS, and hypervisor. Furthermore, by using isolation and cryptography to minimize the attack surface, it improves the security of the processes occurring in TEE. For the secrecy and integrity of data and application code, TEE employs either a hardware or software module, or both. Intel Software Guard Extensions (SGX) [49] is one of the notable TEEs, that are instruction-set architectural extensions that use trusted hardware to offer a secure computing environment. A privacy-preserving international cooperation platform for analysing rare disease genomic data has been proposed in [50]. Intel SGX is used in this study to do secure calculations on dispersed and encrypted genomics data.

2) Secure multiparty computation: Multiparty Computation (MPC) allows distributed computations on encrypted data without the need for decryption. Each data input is split into two or more shares, which are then distributed among several (untrustworthy) parties. Without exposing their inputs to the other parties, all parties follow a procedure and jointly calculate a function on their inputs. They are all given a copy of the final result. It is not necessary in MPC to keep all data from various parties centrally, which would need the use of a trustworthy third party.

MPC has a lower computational cost in comparison to Homomorphic encryption (discussed in the next subsection), but it has a high communication cost. It has to be considered that MPC alone does not provide the output privacy, as the final result may leak some information about the data. Therefore, it can be used in combination with differential privacy. To construct secure MPC protocols, homomorphic encryption [51], garbled circuits [52], linear secret sharing [53], and Oblivious Random Access Machine techniques [54] have been utilised. MPC has been used in conducting queries on genomic data to generate knowledge in [55]. It has also been used in [56] to provide secure genome-wide association analysis.

3) Encryption with special feature: There are several encryption mechanisms with special features such as homomorphic encryption [57], searchable encryption [58], attribute based encryption [59], identity-based encryption (IBE) [60], proxy re-encryption [61], order-preserving encryption (OPE) [62], etc. that can also help the visions. We limit our discussion to the first three methods to provide the readers an overall sense of the potential utility of these techniques.

a) Homomorphic Encryption (HE): It permits calculations (arbitrary functions) to be performed on encrypted data without the need to decrypt it. The data and results, which are both encrypted, would be unknown to the computing environment. As a result, HE makes it possible to do safe computation on an untrustworthy computing platform. HE requires high computations and overhead, which makes it challenging to implement in real-world applications. Some optimisations have been done, but in the general case, it is still insufficient. HE has been used in literature for private computation on genomic data [63], secure testing for genetic diseases on encrypted genomes [64], and achieve genome-wide association study in [65], which securely and privately examines genetic variations and single-nucleotide polymorphisms of genetic data.

b) Searchable Encryption (SE): Classical encryption of data may resolve growing concerns about the protection of outsourced data. It is not simple in reality, though, because the cloud server cannot search the encrypted data directly. Therefore, the user must download all the information and conduct the search after decrypting the data. Searchable Encryption (SE) can solve this problem; it allows the data owner to store data in encrypted form on the cloud while maintaining the server's ability to search through encrypted data. SE has been used to let analysts/clinicians execute queries on genomic data without decrypting it [66]. It has also been studied to protect genomic data when enabling substring search [67], or range query [68] over encrypted genomic data sequences.

c) Attribute-based encryption (ABE): It is a special type of encryption in which the user's secret key and the encrypted data are dependent on the attributes (such as location, job, specialty, etc.). The encrypted data can only be decrypted if the user key's set of attributes matches the ciphertext's set of attributes. It has two types: key-policy attribute-based encryption (KP-ABE) [69] and ciphertext-policy attribute-based encryption (CP-ABE) [70]. CP-ABE has been used along with HE in [71] to provide access control based on the attributes of the users of the system while they request for tests/computation on the genomic data.

4) Blockchain: A distributed public ledger, or blockchain, is a chain of blocks that keeps track of transactions or data in a sequentially growing list. The data included in the blocks are both immutable (no one party has the ability to remove it) and time-stamped. Blockchain facilitates data exchange without relying on a network's compute nodes, and it is not controlled by a central node. The data integrity of blockchain allows the security of networks and systems. Blockchain is utilised in conjunction with homomorphic encryption and secure multiparty computing to provide end-to-end secrecy of genomic data queries [72]. Blockchain is also studied to provide dynamic consent framework [73]. Despite the possibility to increase data security by blockchain data encryption, there is the risk of data leaking from a public blockchain as a result of attacks, such as linkage attacks [74].

We next present a conceptual model of the collaborative analysis of genomic data that summarises our visions.

VI. CONCEPTUAL SYSTEM MODEL

In a practical scenario and based on our visions, the data is stored in a distributed manner. Stored data can be genomic data, medical records, health data that comes from health trackers. The conceptual system model in Fig. 4 considers data is stored on all end devices/institutions, and data are encrypted while stored, processed, or in transit. In this conceptual high-level model, an analyst sends a request (that can be just a query or a model that needs to be trained). The coordinating server (by using the model-to-data method) sends the received model to the data holders, in which the consent has been given and is valid. Training a model can be done using federated learning and split learning techniques based on the computation resources and requirements. The query result can also be generated by using different encryption mechanisms that have special features. These techniques let the queries or computations be performed on encrypted data. The result of each of the requests is sent to the analyst in encrypted form.
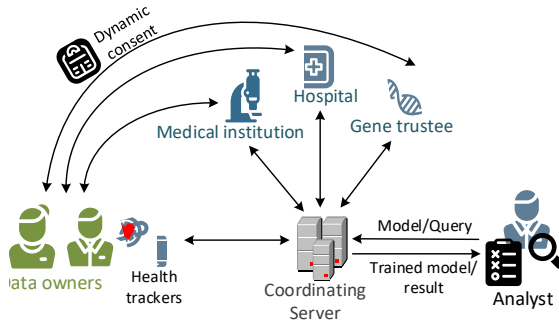

Fig. 4. Conceptual system model

It is worth mentioning that dynamic consent has been considered in this model; there is two-way communication between the data owner and data custodians to let individuals have control over the use of their data. Moreover, considering the computation/query executions/model training are all done on the encrypted data, the secure/private processing and generate knowledge without sharing the data itself are satisfied. Different encryption techniques can be used to support this. This way, by logging all the requests and results, transparency and accountability can be provided. Given the different nature of data stored in different data holders' storage, interconnection and integration can be provided if the requests/models need different data types as input and generate the outputs using this distributed information. It may require some repeats of a process and aggregation at the end to complete the request. Training a model or responding to a query using different distributed datasets, including genomic data and medical records needs further research.

We have designed a model, named PrivGenDB [66], that stores genomic data and other information of data owners in a secure way, and let the researchers/clinicians execute queries on the data in a secure and private way. This way, the data is not shared and just the result of queries is revealed to the authorised researchers/clinicians. PrivGenDB does not address all the visions, e.g., it does not consider dynamic consent, and stores genomic data and other information of data owners centrally. However, the illustrated conceptual system model (Fig. 4) is more general, contains dynamic consent and it recommends having distributed storage. Therefore, a partial solution is proposed in [66] but adapting the full system model presented here and have a solution for all of it is a possible future direction.

VII. OPEN ISSUES

This section categorises and describes the open issues related to the discussed visions.

1) Dynamic consent: Consent may be approached in two ways, as previously mentioned: static consent and dynamic consent. Static consent necessitates approval for all future data uses at the time of data collection, which is usually done on paper. It can't answer problems that develop over time when the environment and requirements change, such as using the data for a different health initiative than the one for which consent was obtained. In this case, dynamic consent is beneficial. The data owner and data custodian have two-way communication, and the owner can update and give various forms of consent or revoke consent. Dynamic consent comes with its own set of problems, including higher implementation costs, consent revocation, and data deletion assurance, as well as the necessity that patients have sufficient technical knowledge and time. Overall, the question of how to automate and manage consent in order to meet regulatory requirements, patient autonomy, cost, and data analytics is still an open issue.

2) Genomic data with medical records: Due to distributed nature of medical records and genomic data storage, it is essential to find a way to combine them for medical and also analysis uses. In the current situation, an individual's medical records are usually stored in hospital storage, and the genomic data is stored in a genome biobank or a gene trustee's storage. For training models, federated learning has already been proposed to develop the solution of training models in a distributed environment. However, there are no works on distributed different datasets that can be used for training a model. Even in the broader perspective, the health data that is collected from smart watches, mobile phones, etc., need also be interconnected with the medical records and genomic data. This is still an open issue that needs to be solved.

3) Combining solutions: Based on the existing challenges and barriers, different mechanisms have been studied and are going to be explored. One of the major concerns would be combining all these solutions on one platform to have all the visions altogether. This is challenging and needs the cooperation of experts from different fields, e.g., in providing genomic data privacy and security requirements, users (clinicians, analysts), biomedical engineers, computer engineers, etc. have to collaborate to be able to combine the solutions. Still, it is an open problem.

4) Decentralisation: In terms of decentralisation, for the better control of individuals over their own genomic data, it can be stored on their own devices. Therefore, anyone who wants to get access has to ask for permission. There are some suggestions of using blockchain and connecting the DNA data to the wallet of individuals, e.g., EncrypGen [75] that utilises blockchain and creates a platform so that users securely store and manage their DNA profiles and earn money by selling access to certain portions of the genome. Decentralisation still has some problems and challenges, such as mobile/personal devices' storage and security, considering emergency situations, and if using blockchain, the blockchain security. Therefore, if considering full control over the data is an important factor, all these challenges must be considered and addressed.

VIII. Conclusion

This article presents our visions of collaborative analysis of genomic data through the lens of moving from regulations and ethics towards utilising different techniques. Considering the sensitivity of genomic data, we discussed regulations and ethics that are currently being used. Then, through examining current detailed problems, our future visions are demonstrated, and some major existing techniques that can support them are listed. To provide an overview of our visions, a conceptual system model is demonstrated and explained. Finally, open issues are elaborated that need further study and research.

References

[1] Satyananda, Vikas and Oshi, Masanori and Endo, Itaru and Takabe, Kazuaki. High BRCA2 gene expression is associated with aggressive and highly proliferative breast cancer. *Annals of surgical oncology.*, pages 1–10, 2021, Springer.

[2] Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nature reviews. Genetics*, 2014. 15(6): 409–21.

[3] Y. Erlich, JB. Williams, D. Glazer, et al. Redefining genomic privacy: trust and empowerment. *PLoS Biol.*, 2014; 12: e1001983.

[4] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):1–44, 2015.

[5] Mohammed Yakubu, Abukari and Chen, Yi-Ping Phoebe. Ensuring privacy and security of genomic data and functionalities. *Briefings in bioinformatics*, Vol. 21, page 511–526, 2020.

[6] Berger, Bonnie and Cho, Hyunghoon. Emerging technologies towards enhancing privacy in genomic data sharing. *BioMed Central*, 2019.

[7] Aziz, Md Momin Al and Sadat, Md Nazmus and Alhadidi, Dima and Wang, Shuang and Jiang, Xiaoqian and Brown, Cheryl L and Mohammed, Noman. Privacy-preserving techniques of genomic data—a survey. *Briefings in bioinformatics*, Vol. 20, page 887–895, 2019.

[8] Hubaux, Jean-Pierre and Katzenbeisser, Stefan and Malin, Bradley. Genomic data privacy and security: Where we stand and where we are heading. *IEEE Security & Privacy*, Vol. 15, pages 10–12, 2017.

[9] Lin, Zhen and Owen, Art B and Altman, Russ B. Genomic research and human subject privacy. *American Association for the Advancement of Science*, 2004.

[10] Li, Jiarui and Conzalez Zarzar, Tomas Benjamin and White, Julie and Indencleef, Karlijne and Hoskens, Hanne and Ortega Castrillon, Alejandra and Nauwelaers, Nele and Zaidi, Arslan and Eller, Ryan and Gunther, Torsten and others. Robust genome-wide ancestry inference for heterogeneous datasets and ancestry facial imaging based on the 1000 Genomes Project. *biorxiv*, 2019.

[11] "BiobankUK." UK Biobank, http://www.ukbiobank.ac.uk, 2018.

[12] "IGRS." The 1000 Genomes Project, https://www.australiangenomics.org.au., 2015.

[13] Botstein, David and Risch, Neil. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*, Vol. 33, pages 228–237, 2003.

[14] Abukari Yakubu and Yi-Ping Chen. Ensuring privacy and security of genomic data and functionalities. *Briefings in Bioinformatics*, 21(2):511–526, 2020.

[15] Logsdon, Glennis A and Vollger, Mitchell R and Eichler, Evan E. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, Vol.21, pages 597–614, 2020.

[16] Bonomi, Luca and Huang, Yingxiang and Ohno-Machado, Lucila. Privacy challenges and research opportunities for genomic data sharing. *Nature genetics*, Vol.52, pages 646–654, 2020.

[17] "O. dictionary." Oxford dictionary, https://en.oxforddictionaries.com/definition/law, (Accessed 30 Sep 2021).

[18] "O. dictionary." Oxford dictionary, https://en.oxforddictionaries.com/definition/ethics, (Accessed 30 Sep 2021).

[19] "The Australian Government, Privacy Act 1988" , Federal Register of Legislation, 2019. Compilation No. 80.

[20] "Genetic Information Nondiscrimination Act (GINA).", Pub. L. 110–233, 122 Stat. 881, (May 21, 2008), 42 U.S.C.

[21] Guide to hipaa privacy rule and compliance. http://www.hipaa-101.com/, (Accessed 30 Sep 2021).

[22] "EU, Regulation (Eu) 2016/679 General Data Protection Regulation", Official Journal of the European Union, 2016.

[23] The GDPR and genomic data - PHG Foundation. https://www.phgfoundation.org/report/the-gdpr-and-genomic-data, (2020).

[24] Baker, Dixie B and Kaye, Jane and Terry, Sharon F. Governance through privacy, fairness, and respect for individuals. *eGEMs, Ubiquity Press*, Vol. 4, 2016.

[25] NHMRC, U.A. ARC, The national statement on ethical conduct in human research 2007. https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018#bl, (updated 2018).

[26] Ethical guidelines for biomedical research on human participants. 2006. http://icmr.nic.in/ethical_guidelines.pdf, (updated 2018).

[27] E. Economic, S. Committee. The ethics of big data: balancing economic benefits and ethical questions of big data in the eu policy context, Report. https://www.eesc.europa.eu/sites/default/files/resources/docs/qe-04-17-306-en-n.pdf, 2017 (Accessed Oct 2021).

[28] AMA, Augmented intelligence in health care, Report. https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf, 2019 (Accessed Oct 2021).

[29] Data Breach 101: Top 5 Reasons it Happens. https://www.whoa.com/data-breach-101-top-5-reasons-it-happens/, (Accessed 30 Sep 2021).

[30] TRENDS IN INFORMATION SECURITY STUDY, Research report. https://www.comptia.org/content/research/trends-in-information-security-study, Mar 2015, (Accessed 30 Sep 2021).

[31] Liu, Vincent and Musen, Mark A and Chou, Timothy. Data breaches of protected health information in the United States. *Jama, American Medical Association*, Vol. 313, pages 1471–1473, 2015.

[32] Data Breaches. https://privacyrights.org/data-breaches, (Accessed 30 Sep 2021).

[33] Human Error a top source of Australian data breaches in 2020. https://www.archtis.com/human-error-a-top-source-of-australian-data-breaches-in-2020/, 16 Mar 2021 (Accessed 30 Sep 2021).

[34] Erlich, Yaniv and Shor, Tal and Pe'er, Itsik and Carmi, Shai. Identity inference of genomic data using long-range familial searches. *Science, American Association for the Advancement of Science*, pages 690–694, 2018.

[35] Sawaya, Sterling and Kenneally, Erin and Nelson, Demetrius and Schumacher, Garrett. Artificial intelligence and the weaponization of genetic data. *Available at SSRN 3635050*, 2020.

[36] Lokesh P Nijhawan, Manthan D Janodia, BS Muddukrishna, Krishna Moorthi Bhat, K Laxminarayana Bairy, N Udupa, and Prashant B Musmade. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research*, 4(3):134, 2013.

[37] Aisling A O'Kane, Helena M Mentis, and Eno Thereska. Non-static nature of patient consent: shifting privacy perspectives in health information sharing. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 553–562, 2013.

[38] A. Goldsteen, S. Garion, S. Nadler, N. Razinkov, Y. Moatti, P. Ta-Shma. Brief announcement: a consent management solution for enterprises.

In *Cyber Security Cryptography and Machine Learning CSCML*, pages 189–192, 2017.

[39] Dwork, Cynthia and McSherry, Frank and Nissim, Kobbi and Smith, Adam. Calibrating noise to sensitivity in private data analysis. *Theory of cryptography conference*, page 265–284, 2006, Springer.

[40] McSherry, Frank and Talwar, Kunal. Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, page 94–103, 2007, IEEE.

[41] Raisaro, Jean Louis and Troncoso-Pastoriza, Juan Ramón and Misbach, Mickaël and Sousa, João Sá and Pradervand, Sylvain and Missiaglia, Edoardo and Michielin, Olivier and Ford, Bryan and Hubaux, Jean-Pierre. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. *IEEE/ACM transactions on computational biology and bioinformatics*, Vol. 16, page 1328–1341, 2018, IEEE.

[42] J. Guinney and J. Saez-Rodriguez, "Alternative models for sharing confidential biomedical data," *Nature biotechnology*, vol. 36, no. 5, pp. 391–392, 2018.

[43] Thapa, Chandra and Camtepe, Seyit. Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Computers in biology and medicine*, Elsevier, pages 6104130, 2020.

[44] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[45] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.

[46] NVIDIA Developer, Nvidia and king's college london debut first privacypreserving federated learning system for medical imaging. https://news.developer.nvidia.com/first-privacy-preserving-federated-learning-system/, 2020.

[47] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.

[48] M. B. Makarious, H. L. Leonard, D. Vitale, H. Iwaki, D. Saffo, L. Sargent, A. Dadu, E. S. Castaño, J. F. Carter, M. Maleknia *et al.*, "Genoml: Automated machine learning for genomics," *arXiv preprint arXiv:2103.03221*, 2021.

[49] V. Costan and S. Devadas, "Intel sgx explained." *IACR Cryptol. ePrint Arch.*, vol. 2016, no. 86, pp. 1–118, 2016.

[50] F. Chen, S. Wang, X. Jiang, S. Ding, Y. Lu, J. Kim, S. C. Sahinalp, C. Shimizu *et al.*, "Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions," *Bioinformatics*, vol. 33, no. 6, pp. 871–878, 2017.

[51] Damgård, Ivan and Pastro, Valerio and Smart, Nigel and Zakarias, Sarah. Multiparty computation from somewhat homomorphic encryption. *Annual Cryptology Conference*, Springer, pages 643–662, 2012.

[52] Yao, Andrew C. Protocols for secure computations. *23rd annual symposium on foundations of computer science (sfcs 1982)*, IEEE, pages 160–164, 1982.

[53] Shamir, Adi. How to share a secret. *Communications of the ACM*, Vol. 22, page 612–613, 1979.

[54] Gentry, Craig and Goldman, Kenny A and Halevi, Shai and Julta, Charanjit and Raykova, Mariana and Wichs, Daniel. Optimizing ORAM and using it efficiently for secure computation. *International Symposium on Privacy Enhancing Technologies Symposium*, page 1–18, 2013, Springer.

[55] M.Z. Hasan, Md.S.R. Mahdi, Md.N. Sadat, and N. Mohammed. Secure count query on encrypted genomic data. *Journal of biomedical informatics*, 81:41–52, 2018.

[56] Cho, Hyunghoon and Wu, David J and Berger, Bonnie. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, Vol. 36, page 547–551, 2018, Nature Publishing Group.

[57] Dowlin, Nathan and Gilad-Bachrach, Ran and Laine, Kim and Lauter, Kristin and Naehrig, Michael and Wernsing, John. Manual for using homomorphic encryption for bioinformatics. *Proceedings of the IEEE*, Vol. 105, page 552–567, 2017, IEEE.

[58] Zhang, Rui and Xue, Rui and Liu, Ling. Searchable encryption for healthcare clouds: a survey. *IEEE Transactions on Services Computing*, Vol. 11, page 978–996, 2017, IEEE.

[59] Li, Ming and Yu, Shucheng and Zheng, Yao and Ren, Kui and Lou, Wenjing. Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption. *IEEE transactions on parallel and distributed systems*, Vol. 24, page 131–143, 2012, IEEE.

[60] Chatterjee, Sanjit and Sarkar, Palash. Identity-based encryption. *Springer Science & Business Media*, 2011.

[61] Bhatia, Tarunpreet and Verma, Anil Kumar and Sharma, Gaurav. Towards a secure incremental proxy re-encryption for e-healthcare data sharing in mobile cloud computing. *Concurrency and Computation: Practice and Experience*, Vol. 32, 2020.

[62] Ayday, Erman and Raisaro, Jean Louis and Hengartner, Urs and Molyneaux, Adam and Hubaux, Jean-Pierre. Privacy-preserving processing of raw genomic data. *Data Privacy Management and Autonomous Spontaneous Security*, pages 133–147, 2013.

[63] Lauter, Kristin and López-Alt, Adriana and Naehrig, Michael. Private computation on encrypted genomic data. In *International Conference on Cryptology and Information Security in Latin America*, pages 3–27, 2014, Springer.

[64] Zhou, Tan Ping and Li, Ning Bo and Yang, Xiao Yuan and Lv, Li Qun and Ding, Yi Tao and Wang, Xu An. Secure testing for genetic diseases on encrypted genomes with homomorphic encryption scheme. In *Security and Communication Networks*, 2018, Hindawi.

[65] J.J. Sim, F.M. Chan, S. Chen, B.H.M. Tan, K.M.M. Aung. Achieving gwas with homomorphic encryption. In *BMC Med. Genom*, 13 (Suppl 7) (2020) 90, 1–12.

[66] S. Jafarbeiki, A. Sakzad, S. Kasra Kermanshahi, R. Gaire, R. Steinfeld, S. Lai, and G. Abraham, "PrivGenDB: Efficient and privacy-preserving query executions over encrypted snp-phenotype database," https://arxiv.org/abs/2104.02890, 2021.

[67] Y. Uchide, N. Kunihiro. Searchable symmetric encryption capable of searching for an arbitrary string. *Security and Communication Networks.*, Vol. 9, No. 12, pages 1726-1736, 2016.

[68] W. Sun, N. Zhang, W. Lou, Y Th. Hou. When gene meets cloud: Enabling scalable and efficient range query on encrypted genomic data. *IEEE INFOCOM.*, pages 1-9, 2017.

[69] Vipul Goyal, Omkant Pandey, Amit Sahai and Brent Waters. Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data. *ACM CCS.*, 2006.

[70] Bethencourt, J.; Sahai, A.; Waters, B.. Ciphertext-Policy Attribute-Based Encryption. *IEEE Symposium on Security and Privacy.*, pages 321–334, 2007.

[71] Namazi, Mina and Eryonucu, Cihan and Ayday, Erman and Pérez-González, Fernando. Dynamic attribute-based privacy-preserving genomic susceptibility testing. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing.*, pages 1467-1474, 2019.

[72] D. Grishin, J. L. Raisaro, J. R. Troncoso-Pastoriza, K. Obbad, K. Quinn, M. Misbach, J. Gollhardt, J. Sa, J. Fellay, G. M. Church *et al.*, "Citizen-centered, auditable and privacy-preserving population genomics," *Nature Computational Science*, vol. 1, no. 3, pp. 192–198, 2021.

[73] N. Mamo, G. M. Martin, M. Desira, B. Ellul, and J.-P. Ebejer, "Dwarna: a blockchain solution for dynamic consent in biobanking," *European Journal of Human Genetics*, vol. 28, no. 5, pp. 609–626, 2020.

[74] C. C. Agbo, Q. H. Mahmoud, and J. M. Eklund, "Blockchain technology in healthcare: a systematic review," in *Healthcare*, vol. 7, no. 2. Multidisciplinary Digital Publishing Institute, 2019, p. 56.

[75] Encrypgen, 2019. https://encrypgen.com/, (accessed October 8, 2021).