2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG) | 979-8-3503-5815-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SEB4SDG60871.2024.10630232

Homomorphic Encryption for Genomics Data Storage on a Federated Cloud: A Mini Review

Philip Ewejobi
Dept of Electrical & Information Engineering, Covenant Applied Informatics & Communication
African Center of Excellence(CApIC-ACE)
Covenanat University
Ota, Nigeria
philip.ewejobipgs@stu.cu.edu.ng

Kennedy Okokpujie
Dept of Electrical & Information Engineering, Covenant Applied Informatics & Communication
African Centre of Excellence(CApIC-ACE)
Covenant University,
Ota, Nigeria
kennedy.okokpujie@covenantuniversity.edu.ng

Emmanuel Adetiba
Dept of Electrical & Information Engineering, Covenant Applied Informatics & Communication
African Centre of Excellence(CApIC-ACE)
Covenant University
Ota, Nigeria
emmanuel.adetiba@covenantuniversity.edu.ng

Babatunde Alao
Dept of Electrical & Information Engineering, Covenant Applied Informatics & Communication
African Center of Excellence(CApIC-ACE)
Covenanat University
Ota, Nigeria
babatunde.alaopgs@stu.cu.edu.ng

***Abstract*—This paper provides an analysis and review of the fundamental aspects of securing genomics data on a federated cloud. It highlights the concept of cloud computing and the security issues associated with it. Furthermore, the concept of genomics, ethical and privacy concerns relating to genomics data, and existing attacks on genomics data. Various cryptographic approaches to data breach the importance of employing homomorphic encryption schemes to safeguard genomics data. It discusses security issues in cloud computing, the concept of federation in cloud computing, the genomics concept and cryptographic approaches for protecting genomics data.**

Keywords—genomics data, federated cloud, homomorphic encryption, cloud security.

I. INTRODUCTION

In the contemporary landscape, the proliferation of cloud-related activities has surged, leading to a substantial expansion in sensitive information. This surge is fuelled by the rapid growth of interconnected devices, laboratory equipment, genomics data, and cloud computing, catalysing advancements in bioinformatics and the economy. However, this intertwined growth also amplifies cybersecurity and information security challenges [1]. Cloud security issues have surfaced due to the presence of unauthorized users, often referred to as intruders, engaging in malicious activities. These intruders primarily seek to access confidential data [2]. The most daunting challenge in cloud data security is the realization that attackers often outpace organizations, exploiting security vulnerabilities overlooked by company employees. Moreover, the swift evolution of new technologies, particularly cloud and mobile, presents additional hurdles. Attackers adeptly adapt to and exploit these technologies, necessitating cybersecurity experts to remain vigilant and anticipate their tactics. Many security measures concentrate on detecting malware and preventing breaches, leading to a reactive approach rather than proactive measures against current and future threats [3].

The risk of cyberattacks on cloud infrastructure escalates as the healthcare and research sectors become increasingly interconnected through technology. External theft involves malicious actors not affiliated with research or healthcare institutions infiltrating sensitive data, such as genomics datasets and patient records, for nefarious purposes like ransom. They may also exploit patients' private data for fraudulent insurance claims or demand ransom payments from healthcare institutions to regain access to patient data systems. Sophisticated malware and phishing techniques are deployed to install malicious programs or pilfer credentials, potentially compromising entire systems [4,5].

Genomic sequencing yields vast amounts of data with immense potential for research and future healthcare applications. Individuals undergoing genetic testing as part of clinical care or research studies often contribute their genomic data along with relevant clinical or phenotypic information [6]. Professional societies widely endorse sharing donated genomic data [7,6,8]. The American College of Medical Genetics and Genomics strongly advocates for widespread genomic data sharing, emphasizing its pivotal role in realizing the full potential of genomic medicine [9,8]. They argue that broad dissemination of genomic information will enhance patient care and expedite the discovery of remedies for genetic diseases. Building vast datasets through genomic data exchange among research studies, often involving tens of thousands of participants, is feasible, reducing costs and participant burdens [10,8].

The manner in which genetic data is stored, shared, and utilized holds significant implications for data contributors, their families, and the broader healthcare system. Public support is

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

crucial for effective data sharing, but obstacles may dissuade individuals from donating their genomic data [11,8]. Individuals often perceive genetic data as distinct from other health information and evaluate the benefits and drawbacks of sharing it differently. Concerns about unintended consequences, privacy, data security, and access by insurance companies and employers have persisted since the introduction of genetic data. The complexity and volume of information in genetic data are expected to increase, exacerbating these concerns. Understanding the criteria individuals consider before donating their genetic data and ensuring a robust cybersecurity framework for data privacy and security are imperative to address these barriers and maximize the potential benefits of genomic medicine.

Researchers have made notable strides in developing cybersecurity tools, such as machine learning intrusion detection systems (IDS), to counter cyberattacks on cloud infrastructure and safeguard the privacy and security of genomics datasets. However, current methods are vulnerable to data poisoning, leakage, and privacy breaches [12]. Hence, exploring encryption techniques within distributed environments, such as federated learning and homomorphic encryption, is essential to fortify genomics data privacy and prevent data leakage in cloud infrastructure. Federation, a technology facilitating multiple servers or databases to operate as a unified entity, promotes communication and collaboration among autonomous, decentralized organizations. Data remains within jurisdictional borders in a federated architecture, while metadata is consolidated and searchable. This approach differs from models where data is centralized through relocation or duplication. Organizations can interconnect their federated architectures to create a federated data platform, enabling users to access and process data across different entities [13]. Full federation involves federating data and computing access across distributed computing and databases, facilitating querying and collaborative data analysis [14,13]. Conversely, homomorphic encryption (HE) enables computations directly on encrypted data, yielding encrypted outputs. Once decrypted, these results can be compared to outcomes obtained from operations on plaintext data. HE is a suitable cryptographic technique for ensuring data security, allowing for the evaluation of data through computations on encrypted data, mitigating concerns about information leakage and enabling the processing of sensitive data by unreliable servers [12].

Given the resilience of federated architecture and homomorphic encryption in safeguarding data privacy and security, this study integrates both technologies to ensure the privacy and security of genomics data in cloud infrastructure.

II. SECURITY ISSUES IN CLOUD COMPUTING

Undoubtedly, cloud computing is a transformative force in how we utilize the internet and share data [15]. Essentially, the main challenge in implementing an effective security scheme arises from the demands placed on cloud environments. Here are some of the threats encountered by such environments:

- **Malicious insiders:** The extent of access and capability for malicious insiders to infiltrate organizations and compromise assets determines the extent of financial and productivity losses. Consumers need to understand the measures taken by providers to detect and protect against malicious insiders. The lack of precise boundaries and the consolidation of services under a single administrative domain increase the risk of gaining complete control over cloud services or extracting sensitive data by malicious actors [15].
- **Account or facility hijacking:** Unauthorized access to user accounts or facilities can lead to significant harm to critical aspects of cloud computing services and compromise the confidentiality, integrity, and availability (CIA) of those services [15].
- **Shared technology issues:** Cloud service providers leverage shared resources in a scalable manner. However, this can create vulnerabilities that attackers can exploit to gain unauthorized access to data. Shared hardware and software resources, such as shared disk partitions and CPU caches in storage servers, can be targets for unauthorized access. Customers should not have access permissions that would allow them to access other tenants' data or network traffic [15].
- **Mistreatment and unethical exploitation of data:** Cloud computing is susceptible to attacks due to its relatively vulnerable registration system, which lacks effective fraud detection capabilities. Therefore, there is a need for robust security solutions to mitigate these risks [15].
- **Insecure interfaces and application programming interfaces (APIs):** Weak interfaces and APIs can give rise to various security concerns within an organization, affecting the confidentiality, integrity, and availability of systems. These interfaces are utilized for tasks such as provisioning, orchestration, management, and monitoring. As new value-added services are developed using these interfaces, the complexity of the interfaces can increase, subsequently adding complexity to the underlying layered API [15].
- **Absence of adequate cloud security standards:** When organizations adopt cloud services, unanswered questions regarding internal security protocols, data storage practices, access permissions to logs, and the extent of information disclosed by vendors in the event of a security breach can lead to significant threats [15].

III. CONCEPT OF FEDERATED CLOUD COMPUTING

In a broad sense, a federated cloud can allow resource sharing and interchange between diverse application domains and consumer groups spanning several administrative domains. Data-sharing services that are used in a variety of situations, such as international scientific collaborations, disaster response operations, supply chain management, or medical information systems, might be considered among these resources. By securely sharing data with specific partners, a secure approach

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

to selective data sharing can facilitate any form of organisational collaboration. The concept of Virtual Organisation (VO), developed within the grid computing community, aimed to achieve this goal by establishing resource-sharing rules and conditions for sharing functional or analytical services [16,17].

Understanding the core principles of the federation is crucial due to its wide applicability and significant impact. These principles are explained in Figure 1 and Figure 2. Figure 1 explains the process of authentication also authorisation in modern frameworks. In this process, a user obtains identification credentials from an Identity Provider (IdP) (1,2). When a service request is initiated by a user from a Service Provider (SP), they need to provide their authentication information (3). The Service Provider then authenticates the user's authentication information with the IdP (4) before proceeding. Once the IdP responds (5), the SP evaluates the access request (6) by considering the user's valid authentication token and associated permissions and properties. Based on this assessment, the SP determines whether to grant or deny the requested service. [17].

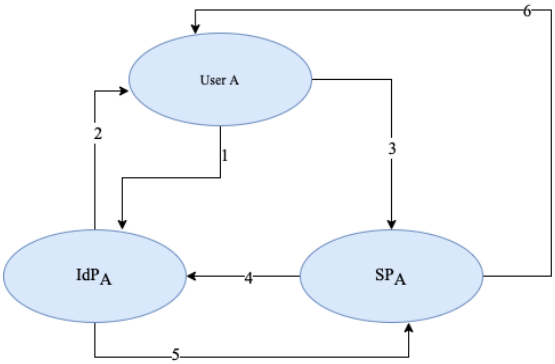

Figure 1: Traditional Authentication and Authorization [17].

To facilitate collaboration among various organisations, it is necessary to establish a similar process among these collaborating entities. This essential need is demonstrated in Figure 2.

Therefore, a federation can be defined as an environment where:

- Users from Organization A have the ability to discover and utilize services offered by Organization B.
- Service Providers within Organization B can authenticate the credentials provided by Organization A and make informed decisions regarding access permissions.

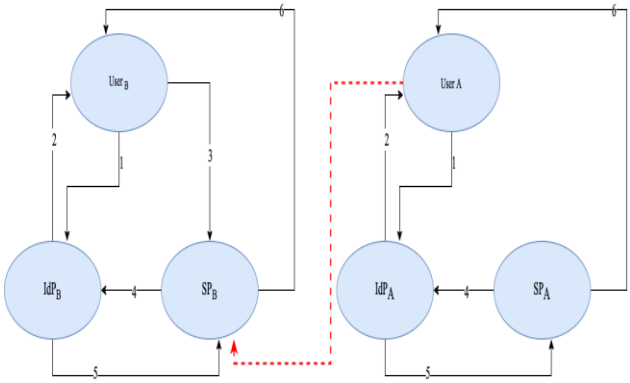

Figure 2: Federated Authentication and Authorization [17].

The upcoming section will delve into the concept of federations, which enable secure resource sharing, including data, platforms, and infrastructure federations, by facilitating authentication and authorization in a decentralized environment.

A. Important Features of Federated Cloud

Based on the understanding of the federation, the fundamental attributes of cloud federation can be identified. These attributes provide a framework for constructing the reference architecture for cloud federation [17]

- A federation is a collaborative and secure virtual environment that is not solely owned by any individual user or organization. It serves as a shared platform for security and cooperation among multiple entities.
- Within a federation, specific users, sites, and organizations collaborate towards shared objectives, with each participating entity having membership and identity credentials associated with them.
- By choosing to share specific resources and metadata, certain entities have the option to become members of a federation. This participation enables the accessibility and discoverability of the shared resources by other federation members.
- Participating members of the federation reach agreements on shared objectives and governance, establishing well-defined access control mechanisms.

B. Federations Serving as Virtual Administrative Domain

The fact that a general federation effectively performs as a Virtual Administrative Domain is crucial to comprehend. Users and resources are handled inside federations in a way that is compatible with other administrative domains. In a federation, however, this domain is virtual, which means it includes a number of elements from numerous sites or organizations. This digital space is not always under the ownership of a single organization but rather serves as a space where participants can come to a consensus on the objectives, goals, and governance of each specific instance of the federation [17].

C. Federation Identity Credentials and Membership

A federation is composed of a group of users who are considered members, based on certain criteria for membership. Each federation can establish its own requirements for membership. Some federations may allow users to join without extensive identity verification or vetting, relying on self-

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

identification. In contrast, alternative federations could possess stringent requirements in place. Certain federations may also impose specific requisites members must adhere to. When joining a federation, members may be required to enter into legal agreements that outline their obligations in supporting the federation's objectives and preventing misuse of shared resources. Additionally, it is important to note that there can be a distinction between individual memberships and organizational memberships, which can significantly impact the governance model of the federation. It can be expressed that these participating entities hold membership in a particular federation since only specific entities opt to collaborate for shared objectives [17].

Federation members may possess identity credentials that are specific to the federation. The definition of a "member" within the federation is based on what the organization decides, and therefore, this choice may also affect the precise format of a member's identity credentials. Additionally, defined procedures and rules control how these credentials are structured and how they are connected to a member's identity inside the original institution where they first joined the federation. [17].

D. Concept of Resource Sharing, Metadata and Discovery in Federation

Although the range of resources, including data and services, that can be shared within a federation is diverse, each federation generally shares specific types of resources aligned with its objectives. It is essential to identify and describe these shared resources using recognized metadata. This requirement highlights the need for semantic interoperability, which can be achieved through standardized schemas and ontologies. Existing work in this field, as well as efforts related to the Internet of Things, can be leveraged by operational federation environments to address this requirement effectively [17].

Once a federation is established, its members will have a need to share and access various resources with one another. There should be a system in place that enables members to learn about the resources and services the federation has to offer in order to assist this. A resource catalog and discovery service are therefore necessary. These catalogue and discovery services can have their implementation specifics and underlying semantics customized to the particular federation. As with cataloged resources, resource discovery policies may be tailored to the federation by taking into account attribute descriptors and user specifications related to each member scanning the catalogue [17].

Members of a federation may, in some situations, work together to develop the discovery policy for different categories of accessible resources. The accomplishment of the federation's goals may depend on how well these efforts are coordinated. On the other hand, in some circumstances, specific resource owners may choose setting their own discovery policies. Sites in the federation have the option to take part by deciding which materials can be found and accessed by other federation members. But it's critical that these regulations follow the resource information, responsibilities, and attributes specified by the federation [17].

When a federation encompasses only a limited and predetermined services that each member provides to others, the resource catalogue and discovery process becomes straight forward. However, in the broader context, resource metadata and service discovery policies become necessary. To assist federation members, it is beneficial to have descriptive database that lists also describes the assets within the federation. This shared repository ensures that vetted information about the resources and services is accessible to all members in a persistent manner. To maintain data integrity, the metadata is cryptographically signed to prevent unauthorized alterations. While the primary objective of a federation is to foster collaboration and resource sharing, it is important to note that resource owners maintain full control over their own resources. Resource owner has the authority to modify their discovery and access policies at their discretion. However, it is crucial for resource owners to have valid justifications for making such unilateral policy changes, as these alterations could potentially have negative consequences for other members of the federation [17].

E. Governance in Federation

The governance of federations is a critical aspect that determines their existence within a broader federation ecosystem. Members who participate in a federation can collectively establish the shared goals and governance framework for their federation. This governance is reflected through policies that govern membership duties and responsibilities, finding resources, and accessing resources [19,17].

To grant or revoke federation membership, a formal process needs to be in place. There must be a system in place for issuing and rescinding memberships if participants cannot simply self-identify and join. This responsibility typically lies with a designated entity called the FedAdmin, which possesses the necessary authorization. This permission could take the form of a position or characteristic given to certain federation members. As part of their role, the FedAdmin is responsible for enforcing identity verification or vetting policies for new members, if any, ensuring that authorized and authenticated users can access the designated resources. If there are conditions in the federation that warrant membership revocation, the FedAdmin is entrusted with executing the revocation process. Additionally, the FedAdmin may have the responsibility of monitoring, detecting, and verifying instances where such conditions arise. Actors who have certain functions or qualities inside the federation ecosystem are involved in federations. The obligations of various members and the actions they can take to determine the policy that will govern the federation's activities are outlined by these roles and qualities. The significance and ramifications of these positions and traits must be understood by every member. A procedure must also be in place for assigning or removing member responsibilities or characteristics [17].

Given that not every member of the federation has access to common resources equally, it becomes necessary to differentiate among the capabilities of different members. This

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

distinction can be achieved through the allocation of diverse user roles or attributes. The entity responsible for granting and revoking member roles or attributes, commonly referred to as the FedAdmin, possesses the necessary authorization to carry out these actions [17].

F. Federated Cloud Architecture

This conceptual model's goal is to identify the fundamental tasks performed by the federation that are significant to a wide range of stakeholders across several application areas. The next sections of this paper provide various approaches for deployment and governance. It is significant to note that there are many different federation solutions that might be used. Some could be rather straightforward, requiring only a portion of the components included in this conceptual model, while others might be complex and call for strong governance mechanisms [17,18].

Figure 3 outlines several components that serve as corresponding counterparts to SP 500-292. These components include Cloud Service Provider, Cloud Service Consumer, Federation Operator, Federation Manager, Federation Carrier, Federation Broker and Federation Auditor

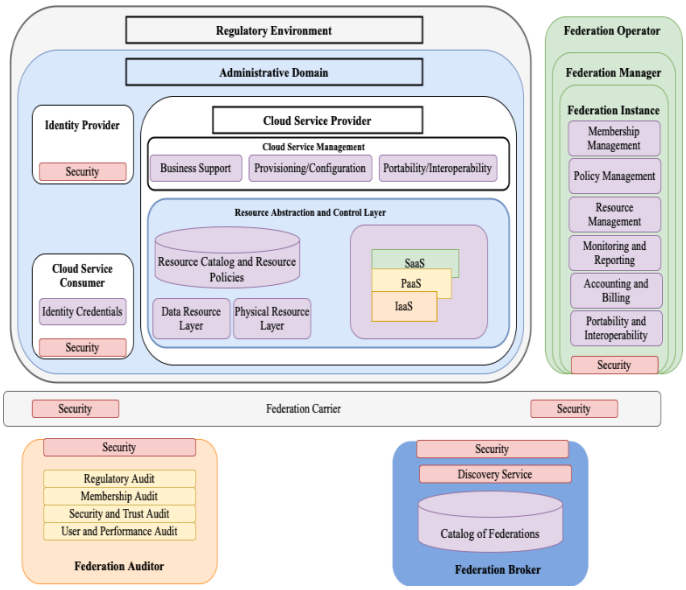

Figure 3: Existing Federated Cloud Architecture [17]

By using this analogy, these components serve as the framework that defines the structure of a federation. While there are numerous similarities, It is crucial to draw attention to a few major modifications and changes to the model that will be covered in more detail. The ideas of Administrative Domains (AD) and Regulatory Environments (RE), which are crucial to this cloud federation paradigm, are two examples. Cloud federations are frequently made up of geographically scattered organizations that function under a variety of legal frameworks that may include numerous international and regional domains [17,18].

The Federation Manager and Federation Operator, two new players who play crucial roles in the management and functioning of the federation, are also introduced under this paradigm. Despite the fact that their duties are separate, their tasks depend on the federation's particular governance architecture [17,18].

G. Security in Federated Cloud Computing

The actors in federated cloud architecture directly address the concerns related to federated identity, authentication, policy, and authorization are integral components. Security negotiations encompass the procedures undertaken to establish a baseline level of trust for interactions among federation members. The goal of the Security Negotiations is To ensure a foundational level of trust for interactions among federation members. The core essence of establishing and managing federated environments lies in creating a secure and collaborative context that can fulfil all the required security needs. Within the framework of a federation, this entails the ability to verify the identity of participants within the federation, determining which resources should be shared within this context, and establishing the mechanism and policy for resource discovery so that only users with permission within the federation can access them, guaranteeing that only authorized users are granted admittance to resources, and ensuring that all interactions maintain the integrity and privacy of information. Specifically, safety has to encompass the requirements for the Federation Manager to detach parts of the federation, allowing for support in cases of repudiation and obsolescence. As we delve into the lifecycle governance requirements of a federation, we will explore these security needs in greater detail [17].

IV. GENOMICS CONCEPT

Biology's field of genomics focuses on examining and studying an organism's entire genome, which is made up of all of its genes. To understand the structure, function, and evolution of genes and how they contribute to the development of an organism's traits, it entails the sequencing, mapping, and analysis of DNA. Our understanding of genetic information has been revolutionized by genomics, which also has several uses in industries including biotechnology, agriculture, and medicine. Genomic research offers the possibility of individualized therapy, targeted treatments, and improvements in genetic engineering and breeding methods. Genomic research is advancing our understanding of the components of life and influencing the direction of future biological research and applications as a result of ongoing technological and computational advances.

A. Ethical and Privacy Concerns of Genomics Data

The potential transformative impact of genomic data is evident; however, the growing accessibility of such data gives rise to numerous unresolved ethical, security and privacy challenges [20]. While various types of health data may contain sensitive patient information, genomic information possesses distinctive and formidable characteristics that raise unique concerns. Genomic data contains information that can be utilized to diagnose and make predictions related to health and behaviour. While these predictions are often probabilistic in nature, they

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

hold substantial diagnostic value and, consequently, raise significant privacy concerns. To address potential discrimination against individuals with genetic markers for certain diseases, laws such as the Genetic Information Non-Discrimination Act (GINA) have been enacted in the United States (U.S.). However, it is important to note that these laws have acknowledged loopholes that may leave the public vulnerable to discrimination from insurance providers, employers, and other entities based on the outcomes of genomic testing. The unwanted disclosure of genomic information can also lead to unintended consequences. For instance, if an individual's genome indicates a predisposition to a specific disease, their family, friends, and employers may exhibit reduced trust in their judgment and decisions [20].

Genomic information possesses a distinct characteristic that sets it apart from other forms of personally identifiable information (PII) or protected health information (PHI). While sensitive data like PIN numbers, home addresses, and bank account numbers are subject to change, genomic information remains inherently sensitive for an extended duration. Unlike transient PII or PHI, genomic information is relatively unalterable and can serve as a lifelong identifier for an individual. Additionally, an individual's genomic information exhibits a significant degree of commonality with that of their blood relatives [21]. This commonality extends not only to existing relatives but also to future generations yet to be born. The genetic composition of a parent will inevitably be interconnected with the genomic makeup of their offspring. As a society, we are just beginning to grapple with the long-term implications of our capacity to uncover the genetic information of individuals.

B. Need for Genomics Data Privacy and Security

Despite the majority agreeing that investigations resulting in the arrest of serial killers benefit society, cases like the Golden State Killer investigation have opened Pandora's box by demonstrating what is possible with unrestricted access to genomic data. According to researchers, the data currently accessible in GEDmatch is considered sufficient for identifying a third cousin or closer in over 90% of the population. When combined with publicly available data and demographic identifiers such as gender, age, or other phenotypic traits, there is enough information to identify a named individual even if their genomic data does not exist in the database and they have never been genotyped. Recent studies have shown that in the United States, by combining searches on public-facing databases with publicly available information, up to 60% of anonymized DNA samples can be uniquely re-identified to a named individual [22,23]. By uploading a small number of specifically designed genomic data files and executing standard queries, malicious actors can retrieve or extract as much as 92% of all genetic markers from user data on publicly accessible databases. This includes the extraction of hundreds of medically sensitive markers [23].

Through the utilisation of these techniques or similar methods employed by law enforcement, malicious actors have the potential to covertly unveil the identities of individuals, including service members and politicians. They can also locate their relatives and unearth compromising information, such as undisclosed children. Since direct-to-consumer (DTC) companies offer genomic data exports in plain text format, third-party services lack the means to authenticate the accuracy of input data. Consequently, malicious actors can upload fabricated data that has been synthetically engineered to determine the target's identity, evade law enforcement, or fabricate relatives and relationships for targeted fraudulent activities. These attacks exploit the distinctive characteristics of genomic information, necessitate minimal computational expertise to execute, and merely require access to standard database and query functionalities available to all users [24].

C. Existing Attack on Genomics Data

Given the highly sensitive nature of genomic data, it comes as no surprise that databases containing such data have been targeted by malicious actors, leading to security breaches (as indicated in Table 1). In May 2019, GEDmatch implemented a change in its data policy, requiring users to explicitly opt-in to facilitating policing agencies' access to their data. However, in November 2019, GEDmatch was served with a warrant from a Florida judge, which demanded complete access to their database for law enforcement purposes. GEDmatch complied with the warrant, inadvertently exposing the data of users who had not opted into sharing their information with law enforcement [25]. Furthermore, in July 2020, a security breach occurred, allowing all GEDmatch profiles to be available for law enforcement matching, disregarding privacy settings [26]. The compromised email addresses obtained from this breach were subsequently targeted in a phishing attack aimed at stealing credentials for the Israeli genealogy website, MyHeritage. Additionally, in 2017, the email addresses and encrypted passwords of 92 million MyHeritage users were illicitly acquired and discovered in a database on an external private server, unrelated to the company.

TABLE 1: REPORTED GENOMIC DATA BREACH

Target	Affected Users	Description
GEDmatch[27]	At the time, GEDmatch had 1.4 million users. It is unknown how many people chose not to share	By overriding a user's opt-in setting, hackers compromised the mechanism that required consent for sharing information with law enforcement. This breach resulted in the exposure of all 1.4 million profiles on GEDmatch, allowing law enforcement to match against these profiles regardless of the users' preferences[27]
dnaLIMS Software[28]	Unknown	By capitalizing on a vulnerability known as CVE-2017-6526 in the web-based sequencing application dnaLIMS, attackers were able to gain unauthorized access to servers and potentially extract DNA sequence hashes through remote control[28]
MyHeritage [29]	Unknown	Users of MyHeritage fell victim to a phishing attack, wherein a counterfeit login form was employed to deceive individuals into disclosing their passwords[29]

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

MyHeritage [30]	92 Million Users	A private server external to the company was discovered to house a database containing email addresses and hashed passwords of all users[30]
Veritas Genetic [31]	The company's official statement regarding the security breach claims that only a small number of customers were impacted	A limited number of customers had their personal information accessed by attackers. Veritas has asserted that the stolen data did not include any personally identifiable health information (PHI).[31]

D. Use Case of Genomic Data

The use cases of genomic data are discussed below:

- **Precision Medicine:** The effectiveness of precision medicine lies in its capacity to enhance the quality of healthcare and improve patient outcomes by guiding clinical decisions towards the most effective treatments [32]. A crucial aspect of precision medicine is developing a comprehensive understanding of genomic information. By integrating data from electronic health records (EHRs) with genotypes and gene sequencing data, significant progress has been made in identifying disease-associated genetic variants, predicting adverse health outcomes, and comprehending responses to therapies [33,34,35]. Successful advancements in precision medicine will steer clinical care away from reactive disease treatment and towards proactive disease prevention and early detection.
- **Medical Research**: To fully achieve the objectives of precision medicine, it is crucial to engage in translational medical research, which heavily relies on the sharing of information and ensuring that researchers can access trustworthy, extensive repositories of electronic health records (EHRs) and genomic data. One important tool utilized by genomics researchers is the genome-wide association study (GWAS), which aims to establish connections between specific genetic variants and particular diseases. In a typical GWAS study, genotypes are collected from multiple individuals in both a study group (consisting of individuals with the disease), and a comparison group (comprising individuals without the disease), focusing on specific loci. Genetic variants that are more frequently observed in the experimental group indicate an association with the disease state. Statistical analyses are conducted to assess the likelihood of a genetic variant being linked to a specific disease, typically necessitating hundreds or thousands of individual genotypes to achieve statistical significance [36]. Recognizing the need for large sample sizes, several data collection initiatives have been initiated to store genomic information and provide access to approved research endeavors. Notable initiatives encompass the All of Us Research Program by the National Institutes of Health (NIH), the Million Veteran Program (MVP) led by the U.S. Department of Veterans Affairs (VA), and the UK Biobank. In the subsequent sub-sections, a brief overview of these initiatives and others will be provided.

E. . Potential Harm of Genomic Data Attack

The primary emphasis of this section is on highlighting the potential adverse outcomes that may arise from a successful attack, as opposed to providing case studies of attacks that have been successfully carried out. Figure 4 illustrates the consequences branch of the taxonomy for genomic data attacks.

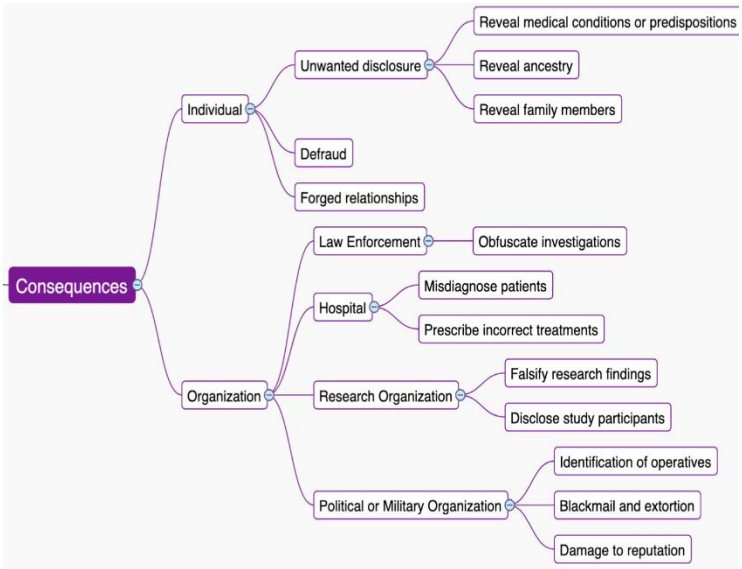

Figure 4:Taxonomy of Consequences of Genomic Data Attack [37,38]

Individual Consequences: The case of the Golden State Killer serves as a tangible illustration of how genomic data can be exploited to reidentify and track down an individual. Although companies like GEDmatch have implemented opt-in programs or restrictions on sharing information with law enforcement, such measures do not eliminate the possibility of a similar attack being conducted by an individual with access to the necessary information. Aside from identifying specific individuals, an adversary could exploit specific genomic markers to reconstruct complete family trees and monitor future generations. An existing example of this genetic surveillance can be found in the Xinjiang region of China, where the Chinese Communist Party employs mitochondrial and Y-DNA data to track the Uyghur Muslim ethnic minority group [37,38]. If an individual's genetic data is compromised, it may also result in the exposure of associated personally identifiable health information (PHI), such as genetic disorders or predispositions to certain diseases that are inherent in their genetic code. There is a possibility that an adversary could exploit these markers to target an individual with the intention of causing harm, such as introducing illicit substances to someone who has a genetic predisposition to addiction. Alternatively, instead of directly harming the individual, a malicious actor may seek to extort the individual for personal gain. This could be achieved by creating a synthetic individual who appears to be closely related to the target or by manipulating their own digital genomic profile to create a desired familial match with the target. Individuals with

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

limited knowledge of genomics and genealogical data would be unlikely to dispute a claim supported by genetic evidence, rendering a significant portion of the population vulnerable to this specific type of attack.

Organization Consequences: Apart from the individual repercussions of genomic data breaches, there are wide-ranging consequences that can impact entire organizations. One of the apparent strategies that adversaries could employ is utilizing the same methodology used to identify the Golden State Killer against law enforcement itself. By assuming that law enforcement would employ a similar approach to identify a suspect in another criminal case, the perpetrator could introduce manipulated data into the system, leading the investigation towards an incorrect family tree branch or falsely incriminating an innocent individual. Although these obfuscation techniques might only temporarily hinder law enforcement, the additional efforts required can result in significant resource expenditure and provide the perpetrator with advanced warning of the investigation, granting them further opportunities to evade law enforcement. If the databases containing specific genomes owned by organizations are compromised, the potential consequences could be severe. As personalized medicine progresses, healthcare systems may seek to possess genetic data from their patients to accurately diagnose and treat various conditions. If a health system's data is compromised by either an insider threat or an external adversary, it could have life-threatening implications for the affected patients. Additionally, it could lead to the disclosure of unwanted information at the individual level, as discussed in the previous section. For instance, two companies are engaged in a competitive race to research a disease with the goal of developing a therapeutic solution. If an adversary manages to manipulate the original data of one of these organizations to create false associations between genomes and related phenotypes, it could lead the institution to invest years of time and resources in investigating a non-existent correlation. This would put them at a considerable disadvantage compared to the unaffected party. It is easy to envision how this scenario could have broader ramifications in both the private and public sectors, highlighting the potential implications of such actions. Furthermore, the impact of genomic data breaches becomes particularly significant when public officials or military personnel are affected. While the individual consequences remain relevant, the societal implications can be far-reaching. In the event that a military officer or prominent politician is compromised by an adversarial foreign entity, there is a risk of subtle manipulation over time, leveraging their compromised status as an insider threat. Such actions could have devastating consequences for national security. If a government or military database is breached, it could grant foreign adversaries access to the data of operatives who depend on forged identities to fulfil their responsibilities [38].

V. CRYPTOGRAPHIC APPROACHES FOR PROTECTING GENOMIC DATA

At its core, cryptographic processes involve a set of encryption and decryption algorithms. Encryption refers to the transformation of data into an unrecognizable form using a cryptographic key and algorithm, with the aim of preventing unauthorized access. The data being transformed is known as plaintext files, while the transformed output generated by the encryption algorithm is known as ciphertext files. On the other hand, decryption algorithms reverse the encryption process by recovering the original plaintext file when provided with the corresponding ciphertext and the correct cryptographic key.

Given the enduring sensitivity of the underlying data, cryptographic methods used to protect genomic data need to ensure long-term security. This requirement cannot be fulfilled by public-key cryptographic algorithms, as their security relies on the assumption of computational hardness. In other words, it is believed that breaking the encryption through brute-force methods would necessitate computational power that is infeasible within a practical timeframe [39]. The effectiveness of widely used encryption methods like the Rivest-Shamir-Adelman (RSA) cypher relies on the computational challenge of factoring large prime integers. However, considering the enduring sensitivity of genomic data, even brute-force attacks that require decades or centuries of computational time may still provide the attacker with relevant and potentially harmful information. Additionally, the emergence of high-performance computing (HPC) tools and powerful graphics processing units (GPUs), along with the anticipated development of quantum computers, poses a threat to the security of algorithms based on computational hardness assumptions. For instance, experts estimate that quantum computers will be capable of breaking the RSA cypher in a matter of hours [40].

A. Concept of Homomorphic Encryption in Genomic Data Security

Homomorphic encryption (HE) allows computations to be performed on encrypted data without requiring decryption. Throughout the entire process, including calculations and analytics, the data remains encrypted, as do the results. Therefore, the system performing the computation does not require access to a secret key. When the outputs are provided to the data owner, they can decrypt them within their own environment using their secret key. Importantly, decrypted data is never exposed outside the user's environment. The outcomes produced by HE schemes are equivalent to those obtained if the calculations were performed on unencrypted data. A common application of HE is in cloud computing environments where data is hosted, and users can conduct computations on it. Data owners only need to encrypt the data once using a public key and store the encrypted data in a cloud repository, such as the UK Biobank. The encryption process satisfies specific mathematical properties, ensuring that computations on the encrypted data correspond to the operations performed on the plaintext data [41].

Within the field of H.E, there are different encryption schemes that offer varying capabilities for performing computations. Partially homomorphic encryption (P.H.E.) enables specific operations, such as addition or multiplication, to be conducted on the encrypted data. On the other hand, fully homomorphic encryption (FHE), often regarded as the ultimate goal in encryption, allows arbitrary computations to be performed on

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

ciphertext, including both addition and multiplication [41]. The security of most F.H.E. schemes rely on the ring learning with errors (R.L.W.E.) computational problem. The security of FHE schemes is built upon the ring learning with errors (RLWE) computational problem, which is related to high-dimensional lattice problems that are currently believed to be resistant to attacks from quantum computers [42]. This is a crucial consideration given the long-term sensitivity of genomic data. In all existing F.H.E. schemes, ciphertexts contain a certain level of noise, which increases during homomorphic computations. If the noise exceeds a certain threshold, with the correct key, the ciphertext remains undecryptable. Partially homomorphic encryption schemes support a restricted set of operations, such as addition and multiplication, until the noise accumulates to problematic levels. This limitation is known as the noise budget. Most somewhat homomorphic schemes are considered feasible within realistic computational constraints. To achieve the unlimited number of operations required for true F.H.E., ciphertexts must undergo constant refreshing during processing to reduce the noise through a computationally intensive process called bootstrapping.

PHE approaches have been applied to various genomics use cases by different research groups. Shimizu et al. developed a string search algorithm using additive P.H.E. to search for sequences of S.N.P.s in databases of aligned genomic sequences [43]. This algorithm enables users to identify the longest matching sequence without revealing the queried sequences. By utilizing standard laptops for issuing queries and executing searches, the total run time for queries over 2,184 genomes ranged from 15 to 133 seconds, depending on the query parameters used. Ayday et al. developed a privacy-preserving approach based on additive P.H.E. to assess the risk of coronary artery disease using S.N.P. profiles [44]. De Cristofaro et al. introduced an approach called Size- and Position-Hiding Private Substring Matching (SPH-PSM). SPH-PSM employs additive P.H.E. to determine the presence of a substring within a genomic sequence file while preserving the privacy of the substring's contents. The primary objective is to enable the owner of encrypted genomic data to verify whether an encrypted substring exists within their genome, providing a binary response of yes or no without exposing the substring itself. The authors conducted benchmark tests on SPH-PSM using substrings of varying lengths. Encrypting the sequence information for an entire genome (approximately three billion base pairs) using a single core on an Intel i7-3770 3.4 GHz quad-core CPU with 16 GB of RAM took 115 hours. Substring searches against the encrypted genome, using a substring size of 1,000 base pairs, required only 0.68ms and scaled linearly with the size of the substring. This search time is considered highly efficient and feasible since most practical genomics cases involving substring matching do not involve lengthy substring templates [45]. Pattern matching of sequence information, which is a crucial task in many genomic applications, poses significant challenges when using fully homomorphic encryption (F.H.E.) due to its complexity. However, Lauter et al. demonstrated the use of FHE in computing algorithms commonly used in genetic association studies, such as the Pearson Goodness-of-Fit and Estimation Maximization algorithms. Their research focused on the single-key homomorphic encryption (HE) schema, where all data are encrypted and decrypted using the same key. For this schema to be applicable in real-world genomics use cases, it would require that all authorized entities depositing encrypted data into a cloud repository be trusted with the key [45].

In the 2018 iDASH competition, there was a track specifically dedicated to secure and parallel genome-wide association studies (GWAS) using FHE [42]. Participants in the competition were tasked with conducting GWAS on encrypted genomic data containing thousands of single nucleotide polymorphisms (SNPs) and hundreds of samples. The winning solution was able to perform an FHE GWAS for 15,000 SNPs within a two-minute timeframe. Notably, all submissions to the competition utilized the open-source Homomorphic Encryption for Arithmetic of Approximate Numbers (HEAAN) framework, indicating a consensus within the community that this framework is well-suited for addressing the numerical optimization challenges present in GWAS [46,42].

B. Sharing genomic data while protecting privacy

Genomic data sharing encompasses two primary approaches: unrestricted access, and limited access, each governed by distinct authorization measures and guidelines. The safety and secrecy of genetic data have been amplified due to various identification attacks. Consequently, there is a reluctance to publicly share genomic data without adequate privacy assurances. However, existing privacy guarantees are insufficient for several reasons, including the numerous assumptions made about adversaries and the presence of different threat models and types of attacks. [46]

C. The confidentiality of query processing and its result

Chen et al. (2016) introduced the concept of location-based skyline queries (LBSQ), which aim to identify Points of Interest (POI) that are non-overlapping by the query position. They proposed a streamlined validation mechanism for LBSQ results obtained from a dubious cloud provider with questionable trustworthiness This mechanism leverages the neighbouring POI relationship to enable effective exploration and validation of the query results.[47] Quan et al. (2018) introduced a method called TOPE (Top Order Preserving Encryption) that allows for top-1 (maximum/minimum) queries to be performed on encrypted data while minimizing information leakage. TOPE achieved this by placing ciphertexts with the top-k values at the beginning of the ciphertext domain [48]. The speed of retrieving the top-k query results from a large, encrypted dataset using the TOPE scheme is nearly as fast as performing the same query on the original unencrypted dataset. Xu et al. (2017) introduced a method for processing aggregated queries on sets of values through top-k query processing. This approach is applied to multiple datasets. To correct the objects of each dataset using range queries, the authors utilize the MG-tree [49]. Ding et al. (2019) made significant contributions by developing several advanced optimization techniques. They introduced a secure technique for range queries, which can serve as foundational components in various representations and protocols [50]. To ensure authentication, a digital signature

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

is appended to each range query. The challenges are addressed through the utilization of encryption schemes commonly employed in homomorphic encryption techniques or by introducing noise to the data, thereby preventing the identification of the strategic plan of the process by third parties [46].

D. Secure genomic data access and storage

Friedrich et al.[51] proposed The TrustStore. The TrustStore refers to a collection of software and web services that enable secure data storage on public cloud infrastructure, such as Amazon S3, OpenStack, or Azure. It ensures data security by fragmenting and encrypting the data on the client machine before transmitting it over the network and storing it in the cloud. Access to the data is restricted to authenticated and authorized users who can decrypt and reassemble it. The TrustStore is specifically designed to support controlled collaboration, allowing users to grant or revoke access to the data they are responsible for. They focused on the integration of TrustStore as a data source within a cloud-based Galaxy environment. The Galaxy environment is utilized for processing (human) genomic data [51].

Wang et al., proposed a framework which is a blockchain-based access control system for Genome-Wide Association Studies (GWAS) using Federated Learning with BFGF (Blockchain-based Federated Learning). Prior to building decentralized models, this system incorporates an Automated Quality Control (AQC) process to ensure the integrity and quality of the training data. To protect the security of users' information, the framework employed an authentication mechanism implemented on the blockchain. This mechanism filtered out malicious attackers, preventing them from accessing sensitive user data right from the start. In order to enhance the efficiency of cloud model training and mitigate various attacks in federated learning, the framework suggests a periodic aggregation method that combines differential privacy mechanisms. This approach aims to speed up the training process on cloud platforms while maintaining the privacy of individual user data. By periodically aggregating the model updates, the system achieves better resistance against multiple attacks. Overall, the blockchain-based framework integrated automated quality control, authentication mechanisms, and a periodic aggregation method with differential privacy to ensure the security and efficiency of GWAS using Federated Learning.

Fujiwara et al.[51] successfully developed a highly efficient and comprehensive system that enables genome-wide data analysis while ensuring controlled secondary data utilization. The model, called the Quantum Secure Cloud System, was designed to operate on a distributed network known as the Distributed Secure Genomic Data (DSGD) analysis system. This network utilized the Tokyo Quantum Key Distribution (QKD) Network, which provided information-theoretically secure communication. [51]

The fundamental building block of the system is a "trusted server" that is built on the quantum secure cloud infrastructure. This server has the capability to deploy and operate various sequencing analysis hardware, including GPUs, FPGAs, and CPU-based software. Through the research done, the paper demonstrated that the DSGD system achieves equivalent data processing speed with and without encryption on the trusted server. This means that the system is fully functional and can be readily implemented in research institutions and medical facilities that perform daily diagnostics outcomes through whole genome sequencing. In summary, the developed system provided a high-performance solution for genome-wide data analysis at a large scale. It ensured secure data access control for both the data custodian and diverse users, and its utilization of quantum secure cloud infrastructure made it an innovative and reliable choice for research and medical institutions. [51]

VI. ACKNOWLEDGMENT

This research work is sponsored by the Covenant Applied Informatics and Communication African Center of Excellence (CApIC-ACE) of Covenant University, Ota, Ogun State, Nigeria and part funding for publication of this paper work was provided through the Google Award for TensorFlow Outreaches in Colleges granted to Prof. Emmanuel Adetiba.

VII. CONCLUSION

Finally, the substantial potential of homomorphic encryption in resolving the privacy and security issues related to the storage of genomics data on federated clouds has been brought to light by this brief assessment. Homomorphic encryption is a viable way to maintain data privacy and facilitate cross-institutional collaborative research by permitting computation on encrypted material without decrypting it. But even with all of its benefits, issues like scalability and processing cost need to be resolved. To enhance homomorphic encryption techniques for the storage of genomics data and to ensure their seamless integration into federated cloud settings, more research and development work is required. In general, the results highlight how crucial it is to keep investigating and applying cutting-edge cryptographic methods in order to guarantee the privacy, accessibility, and integrity of genetic data in cooperative research environments.

VIII. REFERENCES

[1] A. J. Titus, K. E. Hamilton, and M. Holko, "Cyber and information security in the Bioeconomy," Cyberbiosecurity, pp. 17–36, 2023, doi: 10.1007/978-3-031-26034-6_3.

[2] P. Kanagala, "Effective cyber security system to secure optical data based on Deep Learning Approach for Healthcare Application," Optik, vol. 272, p. 170315, 2023, doi: 10.1016/j.ijleo.2022.170315.

[3] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Towards insighting cybersecurity for Healthcare Domains: A comprehensive review of recent practices and Trends," Cyber Security and Applications, vol. 1, p. 100016, 2023, doi: 10.1016/j.csa.2023.100016.

[4] S. T. Argaw et al., "Cybersecurity of hospitals: Discussing the challenges and working towards mitigating the risks," BMC Medical Informatics and

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

Decision Making, vol. 20, no. 1, 2020, doi: 10.1186/s12911-020-01161-7.

[5] S. J. Choi and M. E. Johnson, "The relationship between cybersecurity ratings and the risk of hospital data breaches," Journal of the American Medical Informatics Association, vol. 28, no. 10, pp. 2085–2092, 2021, doi: 10.1093/jamia/ocab142.

[6] A. Middleton et al., "Attitudes of publics who are unwilling to donate DNA data for Research," European Journal of Medical Genetics, vol. 62, no. 5, pp. 316–323, 2019, doi: 10.1016/j.ejmg.2018.11.014.

[7] K. M. Boycott et al., "International cooperation to enable the diagnosis of all rare genetic diseases," The American Journal of Human Genetics, vol. 100, no. 5, pp. 695–705, 2017, doi: 10.1016/j.ajhg.2017.04.003.

[8] F. Lynch et al., "Australian public perspectives on genomic data storage and sharing: Benefits, concerns and access preferences," European Journal of Medical Genetics, vol. 66, no. 1, 2023, doi: 10.1016/j.ejmg.2022.104676.

[9] ACMG Board of Directors, "Laboratory and clinical genomic data sharing is crucial to improving genetic health care: A position statement of the American College of Medical Genetics and Genomics," Genetics in Medicine, vol. 19, no. 7, pp. 721–722, 2017, doi: 10.1038/gim.2016.196.

[10] S. B. Trinidad et al., "Genomic research and wide data sharing: Views of prospective participants," Genetics in Medicine, vol. 12, no. 8, pp. 486–495, 2010, doi: 10.1097/gim.0b013e3181e38f9.

[11] A. Middleton, R. Milne, M. A. Almarri, S. Anwer, J. Atutornu, E. E. Baranova et al., "Global public perceptions of genomic data sharing: What shapes the willingness to donate DNA and health data?" *The American Journal of Human Genetics*, vol. 107, no. 4, pp. 743–752, 2020. [Online]. Available: 10.1016/j.ajhg.2020.08.023

[12] C. Song and X. Shi, "Secure Deep Learning on Genomics Data via a Homomorphic Encrypted Residue Activation Network," *doi:10.1101/2023.01.16.524344*, 2023. [Online]. Available: 10.1101/2023.01.16.524344

[13] M. Alvarellos, H. E. Sheppard, I. Knarston, C. Davison, N. Raine, T. Seeger et al., "Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics," *Frontiers in Genetics*, vol. 13, 2023. [Online]. Available: 10.3389/fgene.2022.1045450

[14] S. Chaterji, J. Koo, N. Li, F. Meyer, A. Grama, and S. Bagchi, "Federation in genomics pipelines: Techniques and challenges," *Briefings in Bioinformatics*, vol. 20, no. 1, pp. 235–244, 2017. [Online]. Available: 10.1093/bib/bbx102

[15] B. Seth and S. Dalal, "Addressing Security in Cloud Federation: A Review," *International Journal of Research in Electronics and Computer Engineering*, vol. 6, no. 3. [Online]. Available: ISSN: 2393-9028 (Print), ISSN: 2348-2281 (Online)

[16] M. Hogan, F. Liu, A. Sokol, and J. Tong, "NIST Cloud Computing Standards Roadmap," *doi:10.6028/nist.sp.500-291v1*, 2011. [Online]. Available: 10.6028/nist.sp.500-291v1

[17] C. A. Lee, R. B. Bohn, and M. Michel, "The NIST Cloud Federation Reference Architecture," *doi:10.6028/nist.sp.500-332*, 2020. [Online]. Available: 10.6028/nist.sp.500-332

[18] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger et al., "NIST Cloud Computing Reference Architecture," *doi:10.6028/nist.sp.500-292*, 2011. [Online]. Available: 10.6028/nist.sp.500-292

[19] J. Kiljander, A. D'elia, F. Morandi, P. Hyttinen, J. Takalo-Mattila, A. Ylisaukko-Oja et al., "Semantic interoperability architecture for Pervasive Computing and internet of things," *IEEE Access*, vol. 2, pp. 856–873, 2014. [Online]. Available: 10.1109/access.2014.2347992

[20] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux et al., "Privacy in the genomic era," *ACM Computing Surveys*, vol. 48, no. 1, pp. 1–44, 2015. [Online]. Available: 10.1145/2767007

[21] L. Bonomi, Y. Huang, and L. Ohno-Machado, "Privacy challenges and research opportunities for Genomic Data Sharing," *Nature Genetics*, vol. 52, no. 7, pp. 646–654, 2020. [Online]. Available: 10.1038/s41588-020-0651-0

[22] Y. Erlich, T. Shor, I. Pe'er, and S. Carmi, "Identity inference of genomic data using long-range familial searches," *Science*, vol. 362, no. 6415, pp. 690–694, 2018. [Online]. Available: 10.1126/science.aau4832

[23] P. Ellenbogen and A. Narayanan, "Identification of Anonymous DNA Using Genealogical Triangulation," *doi:10.1101/531269*, 2019. [Online]. Available: 10.1101/531269

[24] M. D. Edge and G. Coop, "Attacks on genetic privacy via uploads to genealogical databases," *eLife*, vol. 9, 2020. [Online]. Available: 10.7554/elife.51810

[25] K. Hill and H. Murphy, "Retrieved from https://www.nytimes.com/2019/11/05/business/dna-database-search-warrant.html," 2019. [Online]. Available: https://www.nytimes.com/2019/11/05/business/dna-database-search-warrant.html

[26] H. Murphy, "Retrieved from https://www.nytimes.com/2020/08/01/technology/gedmatch-breach-privacy.html," 2020. [Online]. Available: https://www.nytimes.com/2020/08 [27] The New York Times." GEDmatch, a DNA Database, Buckles Under Privacy Concerns. July 19, 2019

[27] The New York Times." GEDmatch, a DNA Database, Buckles Under Privacy Concerns. July 19, 2019

[28] R. Pulivarti et al., "Cybersecurity of Genomic Data," Gaithersburg, MD, National Institute of Standards and Technology, 2023.

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

[29] D. A. Haddon, "Attack Vectors and the Challenge of Preventing Data Theft," in *CYBER SECURITY PRACTITIONER'S GUIDE*, pp. 1-50, 2020.
[30] S. Wertheim, "Auditing for cybersecurity risk," *The CPA Journal*, vol. 89, no. 6, pp. 68-71, 2019.
[31] "Genetics Startup Veritas Rocked by Data Breach," *IT Pro*, Accessed on: Month Day, Year. [Online]. Available: https://www.itpro.com/security/34787/genetics-startup-veritas-rocked-by-data-breach#:~:text=DNAscreening%20company%20Veritas%20Genetics%20has%20suffered%20a%20security,data%2C%20DNA-test%20results%20or%20health%20records%20were%20accessed.
[32] G. S. Ginsburg and K. A. Phillips, "Precision medicine: From science to value," *Health Affairs*, vol. 37, no. 5, pp. 694–701, 2018. [Online]. Available: 10.1377/hlthaff.2017.1624
[33] B. N. Wolford, C. J. Willer, and I. Surakka, "Electronic health records: The Next Wave of complex disease genetics," *Human Molecular Genetics*, vol. 27, no. R1. [Online]. Available: 10.1093/hmg/ddy081
[34] J. Zhao, Q. Feng, P. Wu, R. Lupu, R. A. Wilke, Q. S. Wells, J.C. Denny, and W.-Q. Wei, "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction," *Nature Scientific Reports*. [Online]. Available: doi:10.1101/366682
[35] R. L. Kember, A. K. Merikangas, S. S. Verma, A. Verma, R. Judy, S. M. Damrauer, M. D. Ritchie, D. J. Rader, and M. Bućan, "Polygenic Risk of Psychiatric Disorders Exhibits Cross-trait Associations in Electronic Health Record Data From European Ancestry Individuals," *Biological psychiatry*, vol. 89, no. 3, pp. 236–245. [Online]. Available: https://doi.org/10.1016/j.biopsych.2020.06.026
[36] E. P. Hong and J. W. Park, "Sample size and statistical power calculation in genetic association studies," *Genomics & informatics*, vol. 10, no. 2, pp. 117–122. [Online]. Available: 10.5808/GI.2012.10.2.117
[37] S. Wee, "Retrieved from https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html#:~:text=New%20York%20Times-,China%20Uses%20DNA%20to%20Track%20Its%20People%2C%20With%20the%20Help,system%20of%20surveillance%20and%20control.", 2019. [Online]. Available: https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html#:~:text=New%20York%20Times-,China%20Uses%20DNA%20to%20Track%20Its%20People%2C%20With%20the%20Help,system%20of%20surveillance%20and%20control.
[38] Y. Guo, Z. Xia, W. Cui, C. Chen, X. Jin, and B. Zhu, "Joint genetic analyses of mitochondrial and Y-chromosome molecular markers for a population from Northwest China," *Genes*, vol. 11, no. 5, pp. 564. [Online]. Available: 10.3390/genes11050564
[39] J. Braun, J. Buchmann, C. Mullan, and A. Wiesmaier, "Long term confidentiality: A survey," *Designs, Codes and Cryptography*, vol. 71, no. 3, pp. 459–478, 2012. [Online]. Available: 10.1007/s10623-012-9747-6
[40] L. Gyongyosi and S. Imre, "A survey on quantum computing technology," *Computer Science Review*, vol. 31, pp. 51–71, 2019. [Online]. Available: 10.1016/j.cosrev.2018.11.002
[41] M. A. Will and R. K. Ko, "A guide to homomorphic encryption," *The Cloud Security Ecosystem*, pp. 101–127, 2015. [Online]. Available: 10.1016/b978-0-12-801595-7.00005-7
[42] T. T. Kuo, X. Jiang, H. Tang, X. Wang, T. Bath, D. Bu, L. Wang, A. Harmanci, S. Zhang, D. Zhi et al., "iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching," *BMC medical genomics*, vol. 13, no. Suppl 7, pp. 98. [Online]. Available: https://doi.org/10.1186/s12920-020-0715-0
[43] K. Shimizu, K. Nuida, and G. Rätsch, "Efficient privacy-preserving string search and an application in genomics," *Bioinformatics (Oxford, England)*, vol. 32, no. 11, pp. 1652–1661. [Online]. Available: 10.1093/bioinformatics/btw050
[44] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J.-P. Hubaux, "Privacy-preserving computation of disease risk by using genomic, clinical, and Environmental Data," Retrieved from https://www.usenix.org/conference/healthtech13/workshop-program/presentation/ayday
[45] K. Lauter, A. López-Alt, and M. Naehrig, "Private computation on encrypted genomic data," *Progress in Cryptology - LATINCRYPT 2014*, pp. 3–27. [Online]. Available: 10.
[46] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. In Advances in Cryptology – ASIACRYPT 2017 (pp. 409–437). doi:10.1007/978-3-319-70694-8_15
[47] Chen, Wenxin, Liu Mengjun, Zhang Rui, Yanchao Zhang, Shubo Liu (2016). Secure Outsourced Skyline Query Processing Via Untrusted Cloud Service Providers. In IEEE INFOCOM. Retrieved from https://doi.org/10.1109/INFOCOM.2016.7524509.
[48] Quan, H., Wang, B., Zhang, Y., et al. (2018). Efficient and secure top-k queries with top order-preserving encryption. IEEE Access, 6, 31525-31540. https://doi.org/10.1109/ACCESS.2018.2847307.
[49] Xu, C., Chen, Q., Hu, H., Xu, J., & Hei, X. (2017). Authenticating aggregate queries over set-valued data with confidentiality. IEEE Transactions on Knowledge and Data Engineering, 30(4), 630-644.
[50] Ding, X., Ozturk, E., & Tsudik, G. (2019). Balancing security and privacy in genomic range queries. In Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society (WPES'19) (pp. 106-110). Retrieved from https://doi.org/10.1145/3338498.3358652.
[51] Fujiwara, M., Hashimoto, H., Doi, K., Kujiraoka, M., Tanizawa, Y., Ishida, Y., Sasaki, M., & Nagasaki, M. (2022). Secure secondary utilization system of genomic data using quantum secure cloud. Scientific Reports, 12(1), 1–11. https://doi.org/10.1038/s41598-022-22804

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.

Authorized licensed use limited to: Durban University of Technology. Downloaded on March 12,2025 at 19:45:56 UTC from IEEE Xplore. Restrictions apply.