Asia-Pacific Journal of Science and Technology: Volume: 29. Issue: 03. Article ID.: APST-29-03-03. Review Article

APST

Asia-Pacific Journal of Science and Technology

https://www.tci-thaijo.org/index.php/APST/index

Published by the Research and Graduate Studies,
Khon Kaen University, Thailand

Blockchain empowering genomic data management: a short review

Ramya Patturajan[1,*] and Venkatesh Babu Sakthi Narayanan[2]

[1]Department of Artificial Intelligence and Data Science, PSNA College of Engineering Technology, Dindigul, India
[2]Department of Computer Science & Engineering, Christian College of Engineering and Technology, Dindigul, India
[*]Corresponding author: ramyajun26@psnacet.edu.in

Received 22 March 2023
Revised 21 July 2023
Accepted 28 August 2023

Abstract

Recent advancements in genomic research such as genomic sequencing and editing have given incredible benefits for physicians to deliver accurate and precise treatments. But still, there are a few major challenges such as computational and security issues which could suffocate the performance of a system that deals with genomic data storage, sharing, and analysis. Hence, this paper has been presented with the intention of revealing the potential of emerging technologies to solve the aforementioned bottlenecks with appropriate case studies. To understand the significance of cloud computing in the field of genetic engineering, the exploration has been made to discover the various cloud-based solutions that are developed by various academicians and researchers at various time periods for accomplishing effective genomic data storage, sharing, and analysis. From the detailed study, it has been observed that the cloud computing has a scalable resources and infrastructure to resolve computational complexity of genomic data. But it could be identified that there are no prominent features and methods to prevent security breaches. It necessitates the need for blockchain technology which is being considered an optimal solution for handling genomic data securely. Despite its numerous advantages, blockchain technology still presents some notable technological challenges. These include issues related to scalability, high energy consumption, storage overhead, and potential organizational threats. So, in this article, various blockchain frameworks that have been developed so far for genomic data management are explored and facilitate researchers to understand the potential benefits and implementation challenges of blockchain in genomic data management.

Keywords: Blockchain, Cloud computing, Genomic sequence, Genetic engineering, Big data

1. Introduction

The human genome, which consists of 3 million base pairs, is around 100 GB in size, which is almost equal to 102,400 digital pictures [1]. Genomic analysis involves analyzing variations in the genome of one or more individuals. In clinics, unique genetic information such as a person's DNA can be obtained from biological samples such as hair, skin, blood, and saliva [2]. Upon extraction of the biological sample, it is fed into a sample processing unit located nearby, such as a clinic or institute, as shown in Figure 1. To reduce the sequencing cost, DNA produced by a sample processing unit is sent to a sequencing unit that is operating abroad, especially in developed nations. Sequencing is mostly achieved on high-throughput sequencing platforms like Next Generation Sequencing (NGS), which produces digital output of the sequenced DNA. Thus, the digitalized DNA information becomes the source for bioinformatics analysis. The DNA is in chemical form when samples are collected for the sequencing stage, so physical precautions can provide security. Once DNA information is digitalized, the physical safeguards are insufficient to protect privacy because digital data can be easily copied, modified, and shared.

In the year 2003, the Human Genome Project (HGP) accomplished its first complete genome sequencing after 13 years of work [3]. For over a decade, DNA sequencing followed Moore's law, which was started by the co-founders of Intel. Gordon Moore states that ''the number of transistors that can be placed on an integrated circuit board is increasing exponentially, with a doubling time of roughly 18 months'' [1]. In a simple statement,

computers double in speed and half in size every 18 months. From the year 2007 onwards, DNA sequencing technologies have defeated Moore's law due to the arrival of so-called 'next-generation' sequencing (NGS) technologies. Next-generation sequencing (NGS) and third-generation sequencing (TGS) technologies have been used in clinical and scientific research. Sequencing instruments such as the HiSeq X Five, HiSeq X Ten, and NovaSeq can sequence thousands of samples in parallel and generate tens of terabytes of sequencing data in a single day.

One million genomes will necessitate one million TB of storage, which is equivalent to 1000 petabytes (PB) or one exabyte. The International Data Corporation (IDC) estimated worldwide data to be 0.8 ZB (a trillion GB) in 2009 and will be increased to 40 ZB by 2020 [1]. Genomic data is comparable to other types of big data. More than a dozen genomic data centers use more than 100 PB of data. So, it needs an approach for secure data storage and an efficient tool for analysis, access, and sharing. Storing and sharing genomic data degrades performance in terms of transfer speed, power consumption, data loss and leakage, and server failures.

Many strategies are employed to reduce the size of the data generated. One of the most familiar techniques used to reduce the size of genomic data is delta encoding. The result of the delta encoding file is stored in its most popular file format, such as VCF (Variant Call Format). The concept of delta encoding was initially developed by the 1000 Genomes Project and maintained by the Global Alliance for Genomics and Health (GA4GH). But it comes under computationally intensive methods, which slow down the performance of the system. To overcome the above issues, cloud computing services are integrated with genomics. It solves the problems of storage requirements, computational power, and bandwidth because of its distributed nature.

In general, genomic data sets are typically large and complex, requiring substantial computational resources and infrastructure for processing and analysis. Traditional computing systems may struggle to handle the sheer volume and complexity of genomic data, leading to potential bottlenecks in research and clinical workflows. Hence, this paper aims to explore the potential of emerging technologies for overcoming the computational and security issues in genomic data management. Specifically, the focus is to explore the role of cloud computing for genomic data storage, sharing, and analysis as it is a scalable and resourceful platform. By leveraging cloud-based solutions, researchers and healthcare providers can benefit from flexible and powerful computing capabilities to process and analyze genomic data efficiently. While cloud computing offers scalability and computational power, it falls short in providing robust security features to protect genomic data from breaches. This limitation calls for the integration of blockchain technology as a promising solution. Therefore, this article aims to explore various blockchain frameworks that have been developed for genomic data management. By highlighting the potential benefits of blockchain in genomic data security, researchers in the field of genetic engineering can gain insights into how blockchain technology can enhance the storage, sharing, and analysis of genomic data securely and efficiently. The author in [4] has achieved the security of an IoT-based healthcare system using block-chain through certificate generation and verification, and in [5], the author has come up with an enhanced security solution based on block-chain for the healthcare ecosystem, which is more promising and motivated to incorporate block-chain for genomic data processing. The privacy of data based on blockchain is discussed elaborately in [6], and the application of security for an industrial system based on intelligence [7] enables us to get more information on data security.

The paper begins with the introduction of genomic data and subsequently, a detailed study of various cloud-based solutions which are developed for genomic data management to explore the potential benefits and weaknesses of the cloud are given in Section 2. In Section 3, the effectiveness of integrating blockchain with the cloud is explained with appropriate case studies. Finally, the paper summarizes the findings and highlights of the study as a conclusion in Section 4.

2. Related works

Genomic research is an emerging trend in the healthcare industry [8]. It helps to personalize health care by assisting targeted prevention and treatment plans, resulting in improved clinical effectiveness and outcomes. Moreover, Genomic sequencing is the recent research in the healthcare sector which helps scientists, especially in the COVID-19 pandemic to track the spread of the virus, how the virus is changing, and how those changes will affect public health. Based on the result of this research, the countries are informed to take some important steps such as increasing oxygen supplies, opening more beds, and announcing lockdowns to mitigate the spread of the virus. Increased volume of genome data needs to make use of cloud computing technology (Google cloud storage, Nearline, and Code line) which helps many organizations to collect, store and process huge volume of data in a cost-effective manner.

Genomic research requires enormous processing power and scalability to analyze and store large volumes of datasets. So, the Cloud computing platform is equipped with high-performance computing clusters and the most familiar features of cloud computing such as distributed computing, parallelized programming, and virtualization will provide maximum utilization of hardware which reduces storage and computational cost. Cloud-based

solutions such as Google genomics and Google big query enable rapid analysis of data from thousands of genomes. However, privacy-preserving is still a great challenge for cloud service providers.

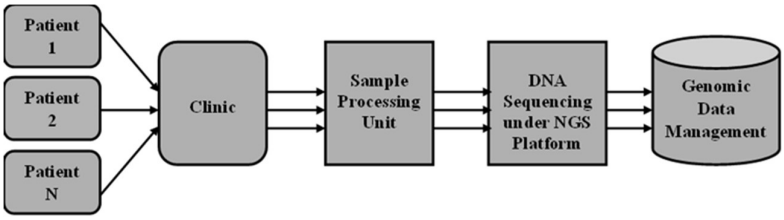

Figure 1 DNA sequencing process.

Genomic data contains some sensitive information about individuals. Leakage of such data leads to severe consequences, including blackmail, discrimination, etc. In 2013, Yahoo's database was cracked, and some security features were added to reset the password. If such a vulnerability happens in the genomic database, the consequences are much worse. As the security expert Bruce Schneier said, if your password is stolen by someone, you can change it. But if your thumbprint has been stolen, you can never get a new one. The failure modes are very different; this statement is opting for genomic data since genomic data is unique to every human being. Without customer knowledge, their data is sold to other companies even after they have signed contract forms. For example, 23 and Me has put its customer data up for sale to another biotechnology company, Genentech, for 60 million dollars. Once an individual's genome is exposed, the corresponding privacy is lost.

To recover from privacy issues, large medical institutions handed over their control of data to third-party stakeholders such as cloud computing companies, which can protect the data and system from security breaches. Many cloud service providers [9], such as Google Cloud Drive, Microsoft One Drive, and Amazon, signed Business Associate Agreements (BAA) and offered HIPAA-compliant services. When users or organizations start to keep their data in the cloud, they lose complete control of that information. So nowadays, many service providers are starting to offer privacy-focused services. For example, SpiderOak, Boxcryptor, and Odrive offer privacy-preserving services through cryptographic encryption. To do encryption, it allows users to choose their passwords. Then cryptographic keys are derived from the given passwords, which are used as symmetric keys for AES encryption. These methods fall under password-based encryption, which has weak security keys and is vulnerable to brute-force attacks. Many years ago, Juels and Ristenpart introduced Honey Encryption (HE), which provides incorrect plaintext when cipher text is decrypted with an incorrect key. It defeats brute-force attacks. But it is inapplicable to more complex structured data, such as genomic data.

Every year, with an increasing number of genetic tests, sequencing becomes more important. An individual's sample is sequenced for a variety of reasons, such as to identify the tendency for diseases, allergies, etc. So, it may lead to some undesired access, which in turn violates privacy. For example, a person named A' possesses a digital record of his DNA sequence and wants to do some genomic testing, which may be conducted by another party, such as a health care provider. Now, the doctor wants to run a query on A's genomic data to see if 'A' is predisposed to a particular disease, but A does not want to share his privacy with that third party. So, he decided to conduct such a test online. It necessitates the need for privacy preservation [2]. Ayday et al. [10] developed a homomorphic encryption-based system in which the storage and processing unit stores genomic data in an encrypted format and authorized institutions perform genomic tests using homomorphic encryption techniques. When private genomic data is queried by publicly known markers, privacy continues to be a problem.

Shortly, government institutions will create genomic databases for preventive health care and personalized medicine. This database contains some additional information, such as location, date of birth, some physical attributes (eyes, skin color, height, and weight), medical history, and biological relationships with other individuals. This database can be owned by some healthcare organizations (database owners) with the help of storage service providers (SSP). It allows researchers to query the database for their research. In this scenario, the researcher may act as an attacker, and they will use some parameters with queries and iterate queries by changing parameters that reveal one or more individuals' identities. So, the privacy of data owners should be guaranteed. To achieve this, a model has been proposed by the authors of [11], which has hardware for securing genomic data processing. Here, it uses tamper-proof hardware. The maximum capacity of a query is the memory size of tamper-proof hardware. However, it is unable to guarantee anonymity when data is taken from the search results.

At the same time, two researchers are working on similar databases on similar topics. Due to financial considerations, they prefer to keep their discoveries a secret from one another before writing or obtaining a patent. In this case, SSP should protect the privacy of data owners and also maintain the confidentiality of queries. To address this issue, a computation technique is presented in [12], where data providers (DP) and data consumers (DC) carry out the computation. Here, DP performs computation on the sensitive regions, and DC performs computation on the rest of the genome. And DC is not given access to sensitive data because sensitive data may

be used to identify people. But in the future, some non-sensitive data will be considered sensitive, from which other sensitive parts can be derived. Moreover, cloud computing has been applied to many IT environments [13] [14] recently in genomics, but it has failed to rectify the following security issues [15]: 1) Anonymity: A specific method or procedure to conceal the published data and essential details, preventing the associated identity of the data's owner. 2) Availability: It provides high client availability and enables users to access their data from any location, at any time. Due to flooding attacks, cloud storage may occasionally be unavailable. 3) Data Leakage and Loss: Data loss results from failed disc drives and data leakage happens as a result of intrusive attacks by attackers. 4) Confidentiality and Integrity issues: There are three basic challenges in cloud storage such as Confidentiality, Availability, and Integrity. Integrity is maintained by protecting the data from unauthorized modification. Cloud data should follow ACID Property to ensure confidentiality and integrity. Since the cloud supports multi-tenant architecture, confidentiality and integrity are violated. 5) Centralized Data Management: Cloud storages are powerful in terms of scalability and usability. But it collects all resources into one data center which leads to a single point of failure.

Contemporary genomic research focuses on solving two important problems in data sharing and access: the privacy of sample providers should be preserved, and sample providers should be able to control access to their data either directly or through trusted third parties such as researchers or doctors with the necessary permissions [16]. Such control can be provided by centralized repositories, but granting and revoking access to data takes long processing times. The centralized architecture is not suitable for assisting the active participation of multiple stakeholders, such as sample providers, doctors, or research groups, in data sharing. It necessitates the need for distributed networks, which distribute approved queries and an encrypted database and facilitate each sample provider's ability to accomplish the data fetch process. Such networks overcome the drawbacks of expense, inefficiency, and security risks associated with a central repository. Blockchain-based platforms are a rising example of distributed networks. A statement stated states, “Each database in such a network is autonomous concerning its data structure, maintains ongoing control of its data, and continuously innovates at its speed” in [3].

Table 1 Major findings of the cloud-based solutions

Cloud solutions	Major findings	Reference
Google genomics and Google big query	• Analysis of data from thousands of genomes. • Privacy preservation is still a great challenge	[8]
Google cloud drive, Microsoft one drive, and Amazon	• Has Business Associate Agreement (BAA) and offers HIPAA-compliant services.	[9]
SpiderOak, Boxcryptor, and Odrive	• Offers privacy preservation service through cryptographic encryption. • Has weak security keys and is vulnerable to brute-force attacks.	[9]
Honey encryption	• Defeats brute force attack. • Inapplicable to more complex structure data such as genomic data.	[10]
Homomorphic encryption-based system	• Storage and processing unit stores genomic data in an encrypted format. • Privacy preserving is still a major challenge.	[10]
Hardware for securing the genomic data	• Uses tamper-proof hardware. • Unable to guarantee anonymity.	[11]
Privacy protection framework	• Provides privacy for data owners and also maintains the confidentiality of queries.	[12]

3. Blockchain in genetic engineering: a survey

Blockchain is a distributed ledger technology that makes it easier for participants to exchange virtual money [17]. It eliminates the need for centralized repositories or trusted third parties (TTP). Traditional, exchanges of electronic currency between users fully depend on mediators such as Banks. Since TTP may act as a malfunction or be compromised by attackers, it makes the financial system insecure, and also TTP leads to a single point of failure. To address these limitations, the first blockchain-based technology has been implemented such as bitcoin cryptocurrency. There are several first-generation blockchain technologies named blockchain−1.0 developed namely Monero, Dash, and Litecoin.

The second-generation blockchain technology named blockchain 2.0 has been developed with some additional features such as contracts with smart attributes. Smart properties are digital assets that allow blockchain technology to govern asset ownership. The rules for how smart properties are governed and managed are encoded

in smart contracts, which are computer codes. Some blockchain−2.0 technologies are NEO, QTUM, Ethereum, and Ethereum Classic. In the beginning, blockchain technologies are apt for financial applications. Now, efforts have been made to make this technology applicable to other non-financial applications. The third-generation blockchain 3.0 is considered a general-purpose technology that applies to many areas such as supply chain management, healthcare, insurance, etc.

Nowadays, blockchain technology came into existence due to its most important aspects, such as decentralization, immutability, and security. As we have seen before, decentralization eliminates the need for third parties, accelerates transactions by removing the lag brought on by a centralized system, and also makes transactions cheaper due to the charging of transactions being removed by the central authority. The advantages of a decentralized blockchain-based security solution are well discussed [18]. The following Figure 2 clearly explains the concept of a centralized vs. decentralized system.

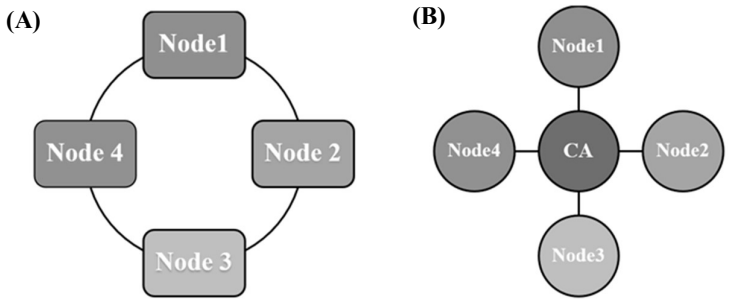

Figure 2 (A) Decentralized approach and (B) Centralized approach.

The centralized system has n number of ledgers, all held by the Centralized Authority (CA). CA has the responsibility to maintain the state of all ledgers. In distributed systems, if there are any discrepancies about the ledger's actual status between two nodes, the CA is then consulted to ascertain the real status of the ledger, but in a decentralized system, there is just one ledger and all nodes have a copy of it with the same degree of access. To maintain the integrity of the ledger, all nodes must agree on the genuine state of the ledger. When all nodes agreed on the true state, it is called consensus.

To impose security measures on the network, cryptographic constructs are employed in the blockchain. Blockchains consist of a collection of participants. Each participant is called a node, and to create a transaction, they must have a pair of keys, such as a public key and a non-sharable key. The publicly available key is used for the identification of the user, and the private key is used for its authentication. Once the transaction is created, it must have the following details: the public key of the user who initiated the transaction, the public key of the transaction's recipient, and the transaction message. All such information can be merged and signed with the user's private key. Then it is broadcast to all the other nodes in the network. This transaction is said to be a proposed transaction.

A block in the network consists of a collection of proposed transactions that are collected within a certain period. All proposed transactions must be valid transaction which ensures that the transaction is legal. That is, it originated from an authorized one (user or node). There are some special nodes in the network called miners which are responsible for validating and ordering the transactions. On receiving the proposed transactions, the miner checks whether a transaction is valid or not. Once the transactions are validated, it is included in the block. Following that, blocks of validated transactions are connected to the earlier blocks to form a blockchain. Ultimately, blockchain is replicated to all nodes in the network, so that all nodes have the same database or ledger of all networked transactions.

The blockchain is created by using another cryptographic mechanism called a hash function. It is a fingerprint algorithm because of its collision-resistant nature. The blockchain is a decentralized ledger that maintains a continuously growing list of records, called blocks, which are linked and secured using cryptography as shown in Figure 3. Each block contains a hash of the previous block, creating a chain of blocks that cannot be altered without breaking the chain. This design enables the detection of any changes made to transactions, providing an immutable record of all transactions.

The immutability of the blockchain is due to the fact that once a transaction is added to a block and the block is chained to the blockchain, it cannot be modified. If any changes are made to the transaction, the corresponding hash of the block will change, breaking the chain to the subsequent blocks in the blockchain. Therefore, the blockchain is said to be an append-only ledger, and any updates to a record can only be made by creating a new record. Additionally, all transactions in the blockchain are time-stamped, providing a detailed log of each transaction's creation, location, and creator. This feature enables auditing and enhances the transparency and

accountability of the blockchain. Overall, the blockchain's decentralized, immutable, and transparent nature makes it a promising technology with numerous potential applications in various fields, including finance, healthcare, and supply chain management.

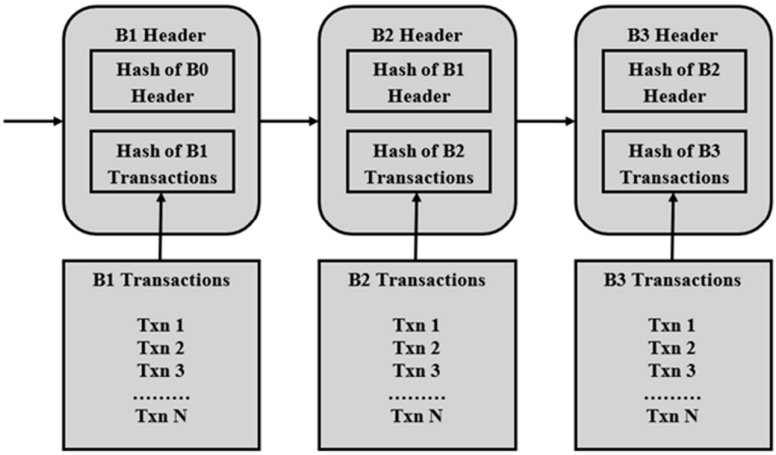

Figure 3 Blockchain network.

3.1 Adaptability of blockchain in genomics

Some salient features of blockchain technology are giving full benefits for healthcare applications which are listed below.

1. Decentralization: Usually Genomic Data Management is a distributed network where different stakeholders are in part such as data owners (patients), healthcare centers, researchers, etc. So blockchain may act as decentralized genomic data management.
2. Openness and transparency*:* Since all data inside the blockchain are duplicated among all nodes in the ecosystem, it facilitates genomic stakeholders especially patients can know how, when, and where their data is used. If one node in the network is compromised, it will not propagate to the ledger due to the replication of the ledger among all nodes in the network.
3. Availability and robustness: The replication of records or ledgers in a blockchain protects genomic data from loss, corruption, and certain security threats. By utilizing blockchain technology, the robustness of the system is enhanced, thereby ensuring the integrity of the data.
4. Data ownership: In the healthcare industry, it is essential for patients to safeguard their data from unauthorized use and to be able to identify any malicious activities. To address these concerns, blockchain technology can be utilized due to its robust cryptographic mechanisms and smart contracts. This allows for a secure and transparent system that ensures patient data is protected and not misused by any stakeholders.
5. Data privacy and security: Blockchain technology employs cryptographic algorithms to safeguard stored data. The decryption of data is restricted to authorized users, thus ensuring data security and privacy. To protect the identities of patients, cryptographic keys are used to pseudonymize their information, allowing health data to be shared among stakeholders without revealing patient identities. Additionally, blockchain incorporates smart contracts, which are programmed to enforce rules that comply with the General Data Protection Regulation (GDPR). These contracts ensure that sensitive patient data cannot be shared or utilized without the explicit consent of the patients.

In recent days, blockchain-based genomic data management has been considered a befitting solution for securely exchanging and storing genomic data. In Xiao-Ling et al. [19] have depicted the elaborated study of a blockchain-based genomic big data platform, namely LifeCODE.ai, to discover the potential of blockchain in secure genomic data management. The LifeCODE.ai was analyzed based on three essential features of genomic data management, such as traceability, secure data sharing, and preserving privacy. On this platform, once an individual has uploaded his genomic data, the data are stored in an encrypted format, and the data owners reserve the right to grant access to any other stakeholders like researchers, hospitals, insurance companies, etc. Unlike typical mechanisms (centralized architecture), LifeCODE.ai enables data owners to lock down access rights, leading to self-data governance. In addition, the data ownership has been determined based on a private key mechanism, and traceability enables owners to monitor data access, which makes all transactions traceable. To perform secure data sharing, LifeCODE.ai incorporates the salient features of blockchain, such as smart contracts to determine allowable transactions and anti-tampering mechanisms to preserve the originality of data. In addition,

it employs a permission blockchain, which has an access control layer that allows certain permissible actions by certain authorized users. To ensure data privacy, all owners' data and associated transactions are securely recorded. Furthermore, the health data are exported to a searchable encrypted repository where the data, which cannot cause any damage to the owner, can only be searched. Finally, a zero-proof mechanism has been incorporated to protect the data owner from revealing any additional information.

As genomic data surrounds more sensitive information, it is essential to develop a strong cryptographic mechanism to protect its privacy, confidentiality, and access control. As of now, several security approaches have been developed that are inadequate for genomic data because of their larger size [12]. To address the security issues that lie in genomic data management, Young-Hoon Park et al. [20] proposed a secure genomic data management system based on blockchain and Local Differential Policy (LDP). In such a system, genomic data are managed in a blockchain environment where a smart contract is employed to ensure integrity and access control. Once the data storage request arrives, the proposed system encrypts the sensitive parts of the data, which reveal the owner's information, or attaches the noise to the sensitive parts before writing them down to the storage, and subsequently a hash calculation is performed to verify the integrity of the data. In addition, blockchain retains access information to control access requests. Moreover, it employs two kinds of storage systems, namely private storage and semi-private storage, to reduce the possibility of gene data leakage. In private storage, encrypted gene data are stored, which is accessed only by authorized internal users who are assumed to be trustworthy users, and in semi-private storage, the noise is generated for the sensitive portions based on LDP, which are irreversible, and then stored, which is accessed by authorized external users who are assumed to be untrustworthy users, such as researchers, insurance companies, etc. Through this, the data owner's privacy is guaranteed in a way that the original data can exist in private storage that is accessed by permissible insiders, and in another case, the data that exits from the system is considered noise data, which is never irreversible. Then, the data owner can trace the usage and transmission of gene data by a decentralized mobile application that is connected to the blockchain.

In Dennis et al. [21] have been involved in developing the blockchain-enabled genomic data sharing and analysis platform, namely Nebula Genomics, with the motivation of reducing the sequencing cost, ensuring data protection, and providing efficient access. It eliminates the need for mediators between data owners and buyers, which enables data buyers to access the data directly from the data owners, which reduces the cost of sequencing. In the Nebula network, data owners are allowed to store their personal genomic and phenotypic data privately and have the potential to lock the access request. The shared data is encrypted using homomorphic encryption, which can be analysed securely using Intel Software Guard Extensions (SGX). While sharing the data, pseudo-anonymity is maintained to conceal information about owners. Moreover, the data buyers are required to be fully transparent about their identities, and all transactions are stored in a ledger in an immutable manner.

The Zenome project is a decentralized and blockchain-based repository for genomic data that provides secure data storage with guaranteed privacy [22]. It allows data users to access the data directly from the owners, and owners can determine the access rights of users. Once data is shared, owners can trace their data usage. The data owners can grant access to third parties with specific constraints, and with the help of private data encryption, disclosure of individual genetic information can be prevented. The distributed environment warrants availability, scalability, and fault tolerance. The Genecoin, or GEN token is a bio-economy currency that is mainly developed to securely exchange genomic data based on the Ethereum blockchain [23]. Here, DNA samples of individuals are encrypted and then stored in the blockchain network. It gets help from third-party service providers to perform genomic sequencing on the collected samples. The system will provide incentives of up to 1,000 Genecoin to the user who recruits another user to this Genecoin network. With the help of supporting software, the extraction and decryption of the genomic data of users from blockchain is possible.

EncrypGen was created as a Hyper Ledger-based blockchain application to securely share and store genomic data [23]. Once the data is uploaded to the network, access permissions is fully controlled by the data owners, and the data sharing is performed in a peer-to-peer manner, which enhances security by eliminating intermediaries. Every user on the network can have a copy of the most recent ledger since blockchain is a decentralized platform. Likewise, while sharing the data, the identity of the owner is concealed as it follows the pseudo-anonymity principle. Here, the users are allowed to upload their data without any constraints, but there may be a chance to upload fake data, which can be resolved by enabling the system to remove unauthorized data. DNATIX is another Ethereum-based blockchain application that is similar to another project in that it uses smart contracts to securely store and transfer genomic data anonymously [24]. It uses its compression algorithms to compress long DNA sequences, thereby compressing 25% of the sequence's size.

Table 2 Effectiveness of blockchain in genomic data management.

Blockchain-based solutions	Major findings	References
LifeCODE.ai	• Genomic big data platform. • Enables data owners to lock the access rights. • Smart contracts to determine the allowable transactions. • anti-tampering mechanism to preserve the originality of data.	[19]
Blockchain and local differential policy	• Smart contract is employed. • Employs two kinds of storage systems to reduce privacy leakage.	[20]
Nebula genomics	• Eliminates the need for mediators between data owners and buyers. • Data owners have the potential to lock the access request. • Pseudo-anonymity is maintained.	[21]
Zenome project	• Owners can determine the access rights of users. • Owners trace their data usage. • Owners grant access to third parties with specific constraints	[22]
Genecoin or GEN token	• Used to securely exchange genomic data based on the Ethereum blockchain.	[23]
EncrypGen	• Securely sharing and storing genomic data. • It follows the pseudo-anonymity principle.	[23]
DNATIX	• Uses smart contracts to store and share data anonymously. • Uses its compression algorithms to compress long DNA sequences.	[24]

3.2 Threats in blockchain

Although there are various benefits of blockchain technology, still faces some technological and organizational threats that need to be considered in future implementation [25].

3.2.1 Scalability issues

While implementing public blockchain applications, there is no restriction on the number of users joining the network which results in scalability issues. Similarly, this issue will arise when wearable devices are connected to blockchain networks to monitor them where the data generated from those devices are at an exponential rate.

3.2.2 High energy consumption

Mostly, blockchains are working based on the Proof of Work Concept which is a mining technique that consumes a greater amount of energy. When the number of users joined in the network is increased, the number of transactions executed per second is also increased results the consumption of more energy.

3.2.3 Storage capacity

As seen earlier, genomic data are larger in size. In the blockchain, every node has a copy of all records, which leads to a shortage of storage capacity for blockchain technology.

3.2.4 Organizational threats

Interoperability is considered the major challenge in implementing blockchain technology due to the lack of trust between healthcare providers and a smaller number of IT professionals with inadequate technical skills for implementing blockchain technology causes terrible consequences.

Table 3 provides a concise summary of the potential advantages and disadvantages of blockchain technology in Genomic Data Management when compared to cloud storage, aimed at facilitating easy comprehension.

Table 3 Comparison of blockchain technology with other methods.

Criteria	Blockchain technology	Centralized database, cloud storage
Potential benefits		
1. Data security and privacy	Highly secure due to its decentralized nature, Immutable and tamper-resistant data, Individual data ownership and control.	Relies on centralized servers, vulnerable to single points of failure, Possibility of data breaches and unauthorized access, and Limited control over data once shared.
2. Data sharing and interoperability	Seamless data sharing across networks, enhanced interoperability between different organizations.	May require complex data-sharing agreements and protocols, Data silos, and limited collaboration potential.
3. Consent management	In transparent and auditable consent management, Individuals have control over data usage.	Consent management may be less transparent and challenging to track, Limited control over data usage once shared.
4. Research reproducibility	A transparent and permanent audit trail, Improved the credibility of research findings.	Lack of transparency and difficulty in reproducing results, Challenges in verifying and reproducing results.
5. Data integrity and auditability	Tamper-proof data records, Enhanced data verification mechanisms.	Auditability depends on the trustworthiness of the centralized system, Susceptible to data manipulation and falsification.
Practical challenges		
1. Energy consumption	High energy consumption due to resource-intensive mechanisms like Proof of Work (PoW)	Cloud storage operates with centralized servers that generally consume less energy for data processing and storage.
2. Storage overhead	Each node in the blockchain stores a complete copy of the entire blockchain, leading to data replication and increased storage overhead	Cloud storage employs data deduplication and management techniques, resulting in lower storage requirements.
3. Scalability issues	In Blockchain, when there are no limits on the number of users joining, scalability becomes a problem as the user base grows larger.	Cloud computing's flexible resource allocation, centralized infrastructure, and load-balancing mechanisms can handle scalability challenges for large user numbers.
4. Organizational threats	Decentralization challenges and regulatory compliance complexities can make blockchain less effective.	Centralized governance and easier regulatory compliance make the cloud more suitable for certain organizational needs.

4. Conclusion

This paper has provided a comprehensive study that facilitates beginners understand the potential of blockchain technology in preserving the privacy of genomic data. Moreover, this paper aims to provide a solution for resolving computational and security issues related to genomic data. As seen earlier, genomic sequence plays a crucial role, particularly in the COVID-19 pandemic. The sequencing process (computational issue) requires more processing power and a scalable platform to collect and analyze such a large volume of genomic data. Several studies have been presented to demonstrate the power of cloud computing in dealing with genomic data, and from this study, it has been found that cloud computing solves computational problems but still suffers from security breaches. It has given me motivation to explore the survey further towards blockchain technology, and numerous genomic projects based on blockchain technology developed recently have been discussed. From this, it has been revealed that there are four potential benefits of blockchain in genomic data management. First, it is decentralized, which enables genomic stakeholders to easily collaborate without the control of central management. Second, the data owner can monitor the data access and can lock the access request. Third, it poses an immutable ledger, and data is kept anonymously, which helps store critical information without anxiety. Finally, data in the blockchain is in an encrypted format and can only be decrypted by the patient’s private key, which prevents attackers from reading the patient's data even after the network is infiltrated. At last, the major threats

that suffocate the smooth implementation of blockchain are also identified, providing valuable insights for future researchers to address and overcome these implementation obstacles in their investigations. In future studies, the significance of quantum computing in solving the computational complexity of genomic data will be revealed.

5. References

[1] O'Driscoll A, Daugelaite J, Sleator RD. Big data, Hadoop and cloud computing in genomics. J Biomed Inform. 2013;46(5):774-781.

[2] Akgün M, Bayrak AO, Ozer B, Sağiroğlu MŞ. Privacy-preserving processing of genomic data: a survey. J Biomed Inform. 2015;56:103-111.

[3] Bourne PE. Confronting the ethical challenges of big data in public health. PLoS Comput Biol. 2015;11(2): e1004073.

[4] Namasudra S, Sharma P, Crespo RG, Shanmuganathan V. Blockchain-based medical certificate generation and verification for IoT-based healthcare systems. IEEE Consum Electron Mag. 2022;12(2):83-93.

[5] Sharma P, Namasudra S, Crespo RG, Parra-Fuente J, Trivedi MC. EHDHE: enhancing security of healthcare documents in IoT-enabled digital healthcare ecosystems using blockchain. Inf Sci. 2023;629:703-718.

[6] Sharma P, Namasudra S, Chilamkurti N, Kim BG, Gonzalez Crespo R. Blockchain-based privacy preservation for IoT-enabled healthcare system. ACM Transactions on Sensor Networks. 2023;19(3):1-7.

[7] Kumari S, Yadav RJ, Namasudra S, Hsu CH. Intelligent deception techniques against adversarial attack on the industrial system. Int J Intell Syst. 2021;36(5):2412-2437.

[8] Google Cloud. Whitepaper handling genomic data in the cloud [Internet]. 2019 [cited 2022 Dec 15]. Available from: https://cloud.google.com/files/genomics-data-wp.pdf.

[9] Huang Z. On secure cloud computing for genomic data: from storage to analysis [thesis]. Lausanne: Swiss Federal Institute of Technology; 2018.

[10] Ayday E, Raisaro JL, McLaren PJ, Fellay J, Hubaux JP. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. The USENIX Workshop on Health Information Technologies (HealthTech '13); 2013 Aug 12; Washington, United States. California: USENIX; 2013. p.1-10.

[11] Canim M, Kantarcioglu M, Malin B. Secure management of biomedical data with cryptographic hardware. IEEE Trans Inf Technol Biomed. 2012;16(1):166-175.

[12] Wang R, Wang XF, Li Z, Tang H, Reiter MK, Dong Z. Privacy-preserving genomic computation through program specialization. The 16th ACM Conference on Computer and Communications Security; 2009 Nov 9-13; Illinois: United States. New York: Association for Computing Machinery; 2009. p. 338-347.

[13] Namasudra S, Roy P. A new table based protocol for data accessing in cloud computing. J Inf Sci Eng. 2017;33(3):585-609.

[14] Namasudra S, Roy P, Balusamy B, Vijayakumar P. Data accessing based on the popularity value for cloud computing. International Conference on Innovations in Information,Embedded and Communication Systems (ICIIECS); 2017 Mar 17-18; Coimbatore: India. New York: IEEE Xplore; 2018. p. 1-6.

[15] Singh S, Jeong YS, Park JH. A survey on cloud computing security: issues, threats, and solutions. J Netw Comput Appl. 2016;75:200-222.

[16] Shabani M. Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems. J Am Med Inform Assoc. 2019;26(1):76-80.

[17] Agbo CC, Mahmoud QH, Eklund JM. Blockchain technology in healthcare: a systematic review. healthcare. 2019;7(2):56.

[18] Namasudra S, Sharma P. Achieving a decentralized and secure cab sharing system using blockchain technology. IEEE Trans Intell Transp Syst. 2022. doi:10.1109/TITS.2022.3186361.

[19] Jin XL, Zhang M, Zhou Z, Yu X. Application of a blockchain platform to manage and secure personal genomic data: a case study of LifeCODE.ai in China. J Med Internet Res. 2019;21(9):e13587.

[20] Park YH, Kim Y, Shim J. Blockchain-based privacy-preserving system for genomic data management using local differential privacy. Electronics. 2021;10(23):3019.

[21] HealthEconomics [Internet]. California: The Company; c1994-2023 [cited 2023 Jan 10]. Blockchain-enabled genomic data sharing and analysis platform. Available from: https://www.healtheconomics.com/white-papers/blockchain-enabled-genomic-data-sharing-and-analysis-platform.

[22] Dedeturk BA, Soran A, Gungor BB. Blockchain for genomics and healthcare: a literature review, current status, classification and open issues. Peer J. 2021;9:e12130.

[23] U.S. Securities and Exchange Commission [Internet]. Washington: The Goverment; c1934-2023 [cited 2023 Jan 23]. Gene-chain version 1.0. Available from: https://www.sec.gov/Archives/edgar/data/1821636/000178043920000057/EXHIBITGbusinessplan.pdf.

[24] DNAtix. The genetics blockchain company, introduces complete genome sequencing pilot and announces token sale [Internet]. 2018 [cited 2023 Jan 15]. Available from: https://www.globenewswire.com/news-release/2018/06/27/1530083/0/en/DNAtix-the-Genetics-Blockchain-Company-Introduces-CompleteGenome-Sequencing-Pilot-and-Announces-Token-Sale.html.

[25] Abu-elezz I, Hassan A, Nazeemudeen A, Househ M, Abd-alrazaq A. The benefits and threats of blockchain technology in healthcare: a scoping review. Int J Med Inform. 2020;142:104246.