Blockchain for Genomics: A Systematic Literature Review

MOHAMMED ALGHAZWI, University of Groningen, Netherlands
FATIH TURKMEN, University of Groningen, Netherlands
JOERI VAN DER VELDE, University Medical Center Groningen, Netherlands
DIMKA KARASTOYANOVA, University of Groningen, Netherlands

Human genomic data carry unique information about an individual and offer unprecedented opportunities for healthcare. The clinical interpretations derived from large genomic datasets can greatly improve healthcare and pave the way for personalized medicine. Sharing genomic datasets, however, pose major challenges, as genomic data is different from traditional medical data, indirectly revealing information about descendants and relatives of the data owner and carrying valid information even after the owner passes away. Therefore, stringent data ownership and control measures are required when dealing with genomic data. In order to provide secure and accountable infrastructure, blockchain technologies offer a promising alternative to traditional distributed systems. Indeed, the research on blockchain-based infrastructures tailored to genomics is on the rise. However, there is a lack of a comprehensive literature review that summarizes the current state-of-the-art methods in the applications of blockchain in genomics. In this paper, we systematically look at the existing work both commercial and academic, and discuss the major opportunities and challenges. Our study is driven by five research questions that we aim to answer in our review. We also present our projections of future research directions which we hope the researchers interested in the area can benefit from.

1 INTRODUCTION

The field of genomics holds great potential in enhancing healthcare. The genomic data produced through technologies such as high-throughput sequencing (HTS) can provide unique health-related information about each individual in a non-invasive manner and is crucial in the advancement of precision medicine [21]. It has been estimated that by 2025, between 100 million and as many as 2 billion human genomes could be sequenced [117]. At the same time as the amount of genomic data is soaring and the genomic technology is advancing, so are the challenges they pose. Who has access to the data [109] and how can they be kept safe? How can data be used and shared responsibly [81, 103] without losing the advantages of sharing for research and (future) patients? “The obligation to confidentiality” must be balanced with “the obligation to share” when it comes to genomic data. Some argue that sharing genomic data is an ethical obligation for those who benefited from sequencing their genome [64]. However, genomic data have certain characteristics that make them fundamentally different from traditional health records: they are long-lived as they carry valid information even after an individual passes away; they indirectly affect descendants and relatives of the data owner; they are large when the whole genome is sequenced (e.g. BAM file size 138 GB [118]).

Moreover, the genomic data are not only used in medical contexts but also in many others including forensics, insurance and pharmaceutical research to name a few. For instance, criminal investigators frequently employ DNA profiles stored in their forensic databases for criminal cases such as rape and murder. Similarly, pharmaceutical companies are investing hundreds of millions to gain access to genomic data towards developing new medicines. For instance, 23andMe [1], a well-known biotechnology company, has been sharing its large genetic database with GlaxoSmithKline [129].

Authors’ addresses: Mohammed Alghazwi, m.a.alghazwi@rug.nl, University of Groningen, Groningen, Netherlands; Fatih Turkmen, f.turkmen@rug.nl, University of Groningen, Groningen, Netherlands; Joeri van der Velde, k.j.van.der.velde@umcg.nl, University Medical Center Groningen, Groningen, Netherlands; Dimka Karastoyanova, d.karastoyanova@rug.nl, University of Groningen, Groningen, Netherlands.

Table 1. Reviews on the use of blockchain in genomics

Paper	Focus
[83, 107]	Potential of blockchain in healthcare with genomics discussed as an example.
[98]	Potential of blockchain in genomics and a use-case proposal.
[110, 124, 125]	Potential of blockchain in genomics.
This paper	Present a comprehensive systematic literature review of the current state-of-the-art methods in applying blockchain in genomics.

While the potential societal impact of the improper use of genetic information is immense, there is a significant public benefit in the adoption of genomic data usage in patient diagnosis, screening and treatment. The main technical issues that have been highlighted in the literature are the storage and sharing [7, 8] of genomic data. More specifically, the need for efficient compression techniques, the lack of harmonized (meta)data, and perhaps more importantly the lack of secure and privacy-preserving technical infrastructures to acquire, process and share genomic data. Another challenge is the lack of common terms and conditions in the metadata, which describes both the data (raw data format and notations) and the access requirements (informed consent, data transfer agreements), decrease the efficiency of discovery and hinders data sharing [128]. In order to maximize the benefits of genomic data, the (meta)data need to conform to FAIR (findable, accessible, interoperable, reusable) principles [29].

1.1 Purpose and Objectives

There is an increasing interest in applying blockchain technology in healthcare. This is evident in the increasing number of articles published each year since the emergence of this technology. For instance, [58] provided an overview of the current trends in using blockchain in healthcare and showed the properties of blockchain that are most commonly used, and [2] classified each work in applying blockchain in healthcare based on the use-case. Blockchain has also been proposed as a candidate approach to address many of the challenges in handling genomic data [98, 110, 124]. The decentralization, immutability, and transparency properties make it an attractive option to solve the sharing and storage issues. In addition, it can be combined with privacy-preserving techniques to provide privacy, traceability, and integrity to the data being managed.

In this review, we do not cover blockchain in general healthcare applications, our focus is on a subset of healthcare applications - genomics. Specifically, we focus on applications and solutions that utilize blockchain technology in managing, sharing, or processing human genomic data. Throughout the paper, we use the term "genomic data" to refer to data used in studying the human genome which includes the scientific study of gene interactions and complex traits/diseases that are caused by a combination of genetic and environmental factors. The use of blockchain in genomics is believed to have great potential as various researchers have pointed out. These works, as shown in Table 1 are focused on the potential and possible benefits of blockchain in genomics, and they list the possible use-cases with few examples works that have been done. For example, a recent paper [124] focused on exploring the opportunities and challenges of DLT in genomics by conducting a ranking-type Delphi study. However, instead of focusing on the potential and possible use-cases, the main objective of this paper is to present the current state-of-the-art methods in applying blockchain in the field of genomics. We look at the wide range of applications, motivations for blockchain, and the different approaches. Finally, we discuss the limitations and future directions, which can serve as a basis for other researchers.

1.2 Review Outline

The review is structured as follows. First, the methodology used to conduct this review is explained in Section 2, then we give an overview of blockchain technology and genomic data storage/sharing in Section 3. In Section 4, we summarize the findings of this review by describing the current trend, application domains, and motivations for using blockchain in genomics. In addition, we discuss the different approaches and techniques used and gives an overview of the challenges faced. Section 5 provides a discussion on our findings and the unexplored opportunities of applying blockchain in genomics.

2 REVIEW METHODOLOGY

The procedure used in this paper is in line with the methodology proposed by Kitchenham et al [67]. We utilize the Systematic Literature Review (SLR) approach as it provides a clear and structured way to search the literature and extract relevant information. Figure 1 represents the methodology followed, which is adapted from [19]. The main three stages are planning the review, performing the review, and documentation.

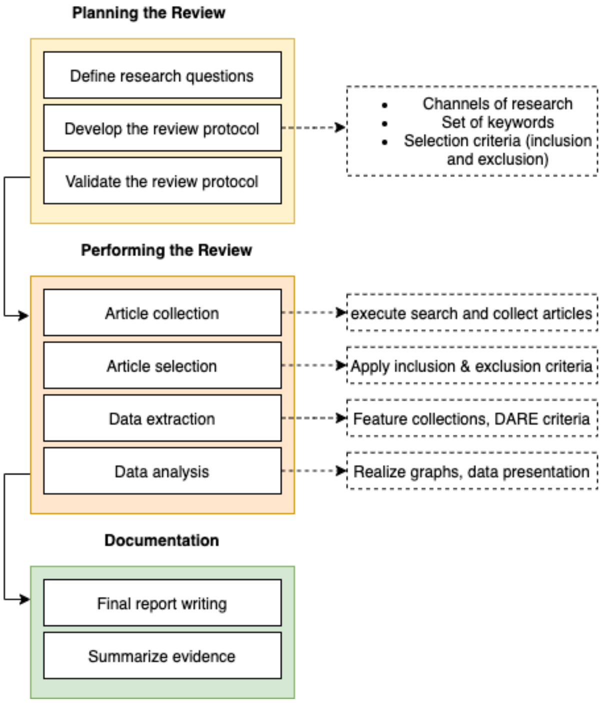
Planning the Review
Define research questions
Develop the review protocol
Validate the review protocol
Channels of research
Set of keywords
Selection criteria (inclusion and exclusion)
Performing the Review
Article collection
Article selection
Data extraction
Data analysis
execute search and collect articles
Apply inclusion & exclusion criteria
Feature collections, DARE criteria
Realize graphs, data presentation
Documentation
Final report writing
Summarize evidence

Fig. 1. Methodology steps adapted from [19].

2.1 Planning the review

The planning for this paper is presented in this section with the first step being the definitions of the research questions which we intended to address. Then, the development of the search protocol is presented to outline the search strategy, and finally, the selection criteria are laid out.

2.1.1 Research Questions. We formulated the main question "What are the application domains, motivations, approaches, and challenges when applying blockchain in genomic applications?" which can be split into the following specific and structured questions that will be addressed in this review:

- RQ1: What are the current research trends for the use of blockchain in genomics?
- RQ2: What are the application scenarios of using blockchain in genomics?
- RQ3: What are the benefits and advantages of using blockchain in these applications as described by the authors?
- RQ4: What are the elements of blockchain technology used in genomic applications? What are the approaches or combinations of technologies used?
- RQ5: What are the challenges and limitations when applying blockchain in genomics? Have these limitations been addressed by the authors? What has been specified for future research?

2.1.2 Search Protocol. The search strategy has been defined and performed after having the previous set of research questions. The selected data sources include both academic and non-academic sources. This is done to cover the wide range of applications and approaches used in both academia and industry.

For academic sources, we collected papers from the following 6 electronic databases:

- Google Scholar,
- IEEE Xplore,
- PubMed,
- Springer SpringerLink,
- Elsevier ScienceDirect,
- ACM Digital Library,

The search for relevant publications in these databases was performed using the query strings defined below:

```
(blockchain OR “block chain” OR "distributed ledger" OR "smart contracts")
AND
(genomic OR genome OR genomics OR genes OR genetic OR genetics)
```

In addition, preprints were collected from (arxiv.org). For non-academic sources, we used the Google search engine to find reports, blogs, and code repositories to select the relevant materials. This is done to find ongoing industry projects that considered the use of blockchain in genomics for commercial purposes. The set of keywords used to find these sources are selected based on the reviewers' background and knowledge related to blockchain and genomic data sharing. These keywords include the following: genomics, genomic data-sharing, blockchain, DLT, smart-contracts. The search was conducted in March 2021 and covered publications in the period 2009 - 2021.

2.1.3 Selection Criteria. The selection criteria (inclusion and exclusion) were defined prior to performing the search strategy in order to eliminate non-relevant sources. The selection criteria are shown in Table 2.

2.2 Performing the review

2.2.1 Article Selection. The initial search in the selected databases resulted in 752 papers. First, a screening of the titles and abstracts was performed and it was aimed to find and exclude duplicate and unrelated articles, which reduced this

Table 2. Selection Criteria

Inclusion Criteria	Exclusion Criteria
1. Original research study (including grey literature).	1. Secondary research, review papers, and non-relevant publications.
2. Original work on the topic of blockchain in genomics.	2. Publications presenting a point of view, magazine publications, interviews, and discussion papers.
3. Publication years in the range between 2009 and 2021.	3. Publications not in English.

number to 61. An additional analysis of the full-text was performed on the remaining articles, and as result, 21 articles were discarded. 40 articles remained and were included in this review as shown in Figure 2.

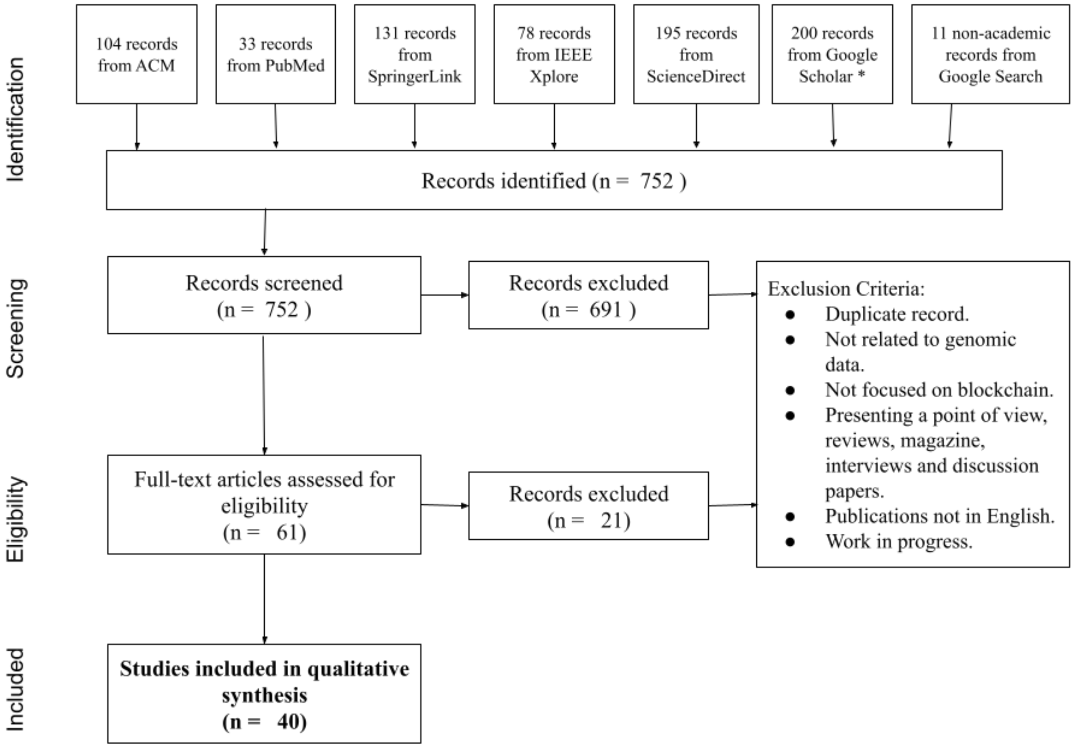
Identification
104 records from ACM
33 records from PubMed
131 records from SpringerLink
78 records from IEEE Xplore
195 records from ScienceDirect
200 records from Google Scholar *
11 non-academic records from Google Search
Records identified (n = 752)
Screening
Records screened (n = 752)
Records excluded (n = 691)
Exclusion Criteria:
● Duplicate record.
● Not related to genomic data.
● Not focused on blockchain.
● Presenting a point of view, reviews, magazine, interviews and discussion papers.
● Publications not in English.
● Work in progress.
Eligibility
Full-text articles assessed for eligibility (n = 61)
Records excluded (n = 21)
Included
Studies included in qualitative synthesis (n = 40)

Fig. 2. Identification and selection process

2.2.2 Data Extraction. We extracted data from each paper to determine the following: (1) the specific blockchain application that researchers focused on. (2) The motivations and advantages of using blockchain. (3) The approach used to implement blockchain for genomic applications, including choices on the blockchain platform, storage, security, and privacy techniques. (4) The limitations and challenges regarding the design and implementation of the proposed application.

3 BACKGROUND

3.1 Genomic Data: An Overview

There are more than 5000 diseases for which a risk level can be calculated by using the genetic information of an individual according to DisGeNET [36]. The genomic data presents an invaluable unique source of information for understanding complex traits and diseases [48]. Traditionally, genomic sequencing (in particular the whole genome sequencing) is considered to be a costly and time-consuming process. Thanks to the developments in genome sequencing technology, today, the time associated with whole genome sequencing is at the level of hours (e.g. one sample in one hour [34]) and the cost is less than 600 Dollars [6]. Large datasets containing genomes and clinical data of individuals are becoming increasingly important to medical experts as the analysis of diverse data contributes to detecting fine-grain biological insights essential to improving public health [14]. As a result, an increasing number of clinicians are including analysis results obtained with these technologies in their day-to-day practices in the context of e.g. personalized medicine. Some of the frequent medical uses of genetic information include diagnostic and predictive DNA testing with the option of integrating polygenic risk scores where an individual's (and her relatives') disposition to certain diseases such as breast cancer is screened through specific genes (e.g. BRCA2).

3.1.1 Genomic data sharing. Interpretation is a key component of genomics research. Individual genomic variants can be interpreted in relation to specific signs or symptoms and multiple genomic variants can be assessed in relation to their collective impact on patients. Genetic (genotype) and clinical data (phenotype) can be combined to determine the best treatment for a patient. The quality of these interpretations is highly dependent on the data they are based on. Research and clinical knowledge sharing is essential to enable the refinement of interpretations.

While policies and laws, which differ from country to country, allow the exchange of genomic data under certain conditions, genomics researchers have experienced how difficult and cumbersome the process is [77]. Another issue is the participation in genomics research which is currently low considering the requirements in genomics research. For instance, the number of participants in genome-wide association studies (GWAS) can reach over 1 million [89].

Current genomic data sharing methods depend on the level of privacy (and the task at hand) required as some parts of the genomic data are private while some others are not considered private. For instance, somatic variants in the human genome are not considered private as they cannot identify specific individuals or families. On the other hand, germ-line variants are unique to each person, and therefore, they require privacy protection. There are various genomic data exchange platforms that give researchers the ability to share genomic data publicly in order to advance the research in this field. Large organisations such as Clinical Genome (ClinGen) [24] and the Global Alliance for Genomics and Health (GA4GH) [46] have started the development of reliable resources to systematically define and interpret all human variation through broad data-sharing efforts. There are also large scale European efforts to promote/coordinate the cross-border collection, storage and sharing of human genome data in a secure way, e.g. Beyond 1 Million Genome (B1MG) [126]. One of the proposed solution is the Genomic beacon project initiated by GA4GH. The genomic beacons ease the process of genomic data sharing through the use of web services by answering queries about the presence of a specific allele in a genome. Institutions can launch their own beacons and connect to the project. There are currently over 100 beacons [44]. In general, beacons aim to respect the data privacy by allowing the institutions to define their own access restrictions and authorization schemes. However recently, privacy researchers have pointed out that the beacons actually have privacy issues and the individuals can be identified from them even if a data anonymization technique is applied to the data [111].

3.1.2 Privacy of genomic data. In addition to being highly valuable, genomic data is highly sensitive as it may reveal information about an individual and/or his/her family. The susceptibility to certain diseases, ancestral traits of an individual and response to a drug are just a few of the use cases demonstrating the private nature of one's genetic information. The privacy concerns around the genomic data are among the main reasons that limit its wide spread use. In addition to (HIPAA) [96] and General Data Protection Regulation (GDPR) [25], there are tailored regulations to address these concerns such as Genetic Information and Nondiscrimination Act (GINA) [26].

The conflicting need for genomic data to be both shared and private requires the use of privacy-preserving techniques which allow processing of the data while preserving privacy. There is a plethora of research on the topic of privacy-preserving processing of genomic data that propose the use of privacy-enhancing technologies such as Homomorphic Encryption [9, 38] and secret sharing [114, 134] among others. We refer the reader to the surveys on the topic such as [4] for a more systematic presentation.

The size of genomic data that ranges between 30-200GB, is one of the main obstacles for the application of privacy-preserving techniques to genomic data processing. There are proposals to substantially reduce the storage footprint of genomic data as PetaGene's proposal on compressing NGS datasets in FASTQ and BAM format to provide on average 60% reduction while preserving the genotyping accuracy [102]. However, with the large scale collection efforts at the horizon such as B1MG [126], an explosion of genomic data is expected. Furthermore, the large size of genomic data makes it difficult to store locally and therefore cloud databases might be used which introduces other privacy and security issues. In addition, the high-performance computation required for genomic data processing makes it difficult to perform locally as well. Finally, there are various security and privacy attacks on genomic data on current genomic data sharing platforms as listed in [10]. This opens the field to alternative approaches that follow the laws and regulations around privacy and at the same time provide utility.

3.2 Overview of Blockchain

The detailed technical foundation of blockchain technology is outside the scope of this paper. However, it is important to shed light on some blockchain concepts, features, and terminologies that will assist the understanding of how blockchain is applied to solve problems in handling genomic data. For an extensive treatment, we refer the reader to other articles such as Kolb et al. [68] and Zhang et al. [133].

Blockchain is the innovative technology behind Bitcoin, the first open-source decentralized digital currency system. The initial design and implementation were done by an unknown entity named Satoshi Nakamoto in 2008/2009 [92]. Blockchain stores and verifies transactions on a ledger that is distributed to all nodes in a peer-to-peer (P2P) network. The transactions are organized into blocks which are protected by a combination of cryptographic techniques to ensure the integrity of the recorded transactions. A consensus protocol is then followed to validate the blocks and the blocks that are successfully validated are added to the growing chain of blocks. Although blockchain technology and Distributed Ledger Technology (DLT) are closely related, there is a difference. A distributed ledger is a ledger or a database that is spread across the nodes in the network and maintained by a group of peers, rather than a central agency. Blockchain is an implementation of DLT and unlike a database, it consists of a chain of blocks. These data blocks are unique data structures that distinguish blockchains from other DLT types. Other implementations of DLT include Hashgraph [11] and Directed Acyclic Graph (DAG) [140].

3.2.1 Blockchain types. Blockchain can be divided into a few distinct types, which have their own characteristics, and directly reflect the network behavior. These types of blockchain can be classified into the following [120]:

(1) Public Blockchain (public & permissionless): a permissionless ledger in which any anonymous node can join the network, and no trust requirement is enforced by the network members. The transactions are publicly broadcasted to all the nodes. Any node in the network can participate in the consensus mechanisms to validate the blocks. An example of this type of network is the Bitcoin blockchain [92].
(2) Private Blockchain (private & permissioned): a ledger that is managed by a single entity, and permission is required before any node can join the network. The access control mechanism provides a higher degree of privacy to the content of the blockchain transactions. Additionally, this type of blockchain provides higher performance in terms of block confirmations. An example of a platform of this type is MultiChain [45].
(3) Consortium/Hybrid Blockchain (public & permissioned): a ledger that is managed by a pre-selected group of nodes. Similar to private blockchains, nodes require permission to join the network. Validating blocks and transactions is done when a chosen set of nodes reach a consensus. The exact process depends on the pre-established rules of the consensus mechanism. An example of a platform of this type is Hyperledger [18].

3.2.2 Consensus mechanisms. Consensus in blockchain is the process to validate the blocks and their contents (transactions and code) in order to add them to the blockchain. This essentially solves the problem of allowing multiple parties that do not necessarily trust each other to agree on the state of a shared ledger. Consensus protocols are essential for the reliability of the blockchain. Proof-of-work (PoW) in bitcoin was the first used consensus protocol in blockchain. This consensus mechanism is comparable to a competition where nodes (miners) try to solve the same puzzle (preimage to a hash function) to validate the transactions and generate a new block. The node that provides the correct solution to the proof-of-work receives a reward in the form of cryptocurrency such as Bitcoin. The newly generated block is then broadcasted to all peers in the network and that block gets connected to the existing blockchain. PoW has been criticized for its energy waste and slow block confirmation. Various protocols have been developed to overcome some of the limitations in PoW such as Practical Byzantine Fault Tolerance (PBFT) and Proof-of-stake (PoS). For a more detailed description of blockchain consensus protocols, we refer the reader to Xiao et al. [131].

3.2.3 Smart Contracts. Smart contracts have emerged recently on blockchain due to the popularity of the Ethereum platform. However, the concept of smart contracts dates back to 1997 and was proposed by Nick Szabo [122]. The concept has evolved since then, but the main objective is to allow a smart contract program to run in a decentralized network and modify the state of the system in an automated, trusted, and verifiable way without intermediaries. Blockchain has made it possible to implement this concept and use it in different settings including finance, Identity Management, and healthcare [66]. There is a variety of blockchain platforms that allow executing smart contracts using a number of programming languages and one of the most commonly used languages is Solidity. Briefly, writing a smart contract involves establishing a set of requirements and instructions that are automatically executed once these requirements are met. In contrast to written contracts, smart contracts are executed automatically, are publicly verifiable, and do not require any intermediaries.

3.2.4 Security and Privacy in Blockchain. The security and privacy features of blockchain rely on the use of a number of cryptographic techniques. Some of these techniques were leveraged by the original bitcoin blockchain design, while others were added to subsequent blockchain implementations to enhance security and privacy. The basic security and privacy techniques utilized in the bitcoin blockchain ensure that the system meets the security and privacy related requirements of online transactions and prevents known vulnerabilities. These requirements include: consistency of the ledger, the integrity of the transactions/data, availability of the system, confidentiality of transactions, and users' anonymity [133].

Blockchain ensures the consistency of the ledger across multiple nodes through the use of a consensus mechanism discussed previously. The integrity of online transactions is essential and the underlying system used to facilitate them must be secured against malicious tampering of the data. Blockchain transactions are resistant to tampering from both miners that confirm those transactions, and the external attackers that try to manipulate blockchain transactions. Using cryptographic hashing and digital signature, any modification on the transaction data would be detected by checking the validity of the digital signature. In addition, tampering with blockchain transactions requires altering the data stored in all blockchain nodes since the ledger of transactions is stored in all nodes in the network. Attacks such as distributed-denial-of-service (DDoS) are not feasible because of the highly decentralized nature of the blockchain network. It is important to note that the more blockchain nodes there are in the network the higher the resistance to the attacks on availability.

Regarding the privacy aspect, blockchain provides pseudonymity through the use of public key infrastructure (PKI). The nodes or users are identified by their public addresses rather than their real identities. However, this fails to provide full anonymity as there is a risk of linking the public address to real identities by observing the interactions between different parties [133]. Another privacy limitation of the original blockchain implementation is that transactions and their data are publicly visible. Ensuring the confidentiality of transactions and smart contract data expands the possible applications of blockchain to include those that handle private and sensitive information.

To overcome the aforementioned limitations, additional security and privacy techniques have been proposed such as mixing, anonymous signatures, and Zero-Knowledge Proofs (ZKP) [15, 133]. Mixing services have been proposed as a solution to provide unlinkability to transactions [17, 87, 108]. The mixers swap users' coins and prevent tracing the movement of coins in the blockchain, thus provide unlinkability. Anonymous signatures are another privacy cryptographic technique employed by certain blockchain applications to hide the identity of the signer. Anonymous signatures schemes such as group signatures [22] and ring signatures [105] conceal the identity of the signer among a group of users that signed a transaction/message. A further discussion on the additional privacy techniques that can be combined with blockchain is provided in Section 5.3.

4 RESULTS

In this section, we present the outcomes of the review. The aim is to answer the research questions defined in Section 2. We first show the distribution of publications per year. Then, we list and classify each paper into sub-categories based on the application domain, and present our analysis on the motivation of using blockchain as described in the papers. We also present the methods and approaches employed in these papers. Finally, we list the open issues and challenges that we observed from the papers.

4.1 Current Trend in Genomic Blockchain Research

To address RQ1, we analyzed the yearly trend in publications relating to the use of blockchain in genomics. This trend can be seen in Figure 3, which shows an increasing interest in this topic. The interest started with a lot of commercial applications, but with time more academic research followed. Based on our findings, the first use of blockchain in genomics appeared in a commercial application, Genecoin [47], which started in 2014. The number of papers increased over the years with 18 papers in 2020 alone. The quick upward trajectory seen in Figure 3 is expected since blockchain is a relatively new technology that was introduced in 2009 and the implications of its use (e.g. for scalability and security/privacy) are just being studied in a non-cryptocurrency context.

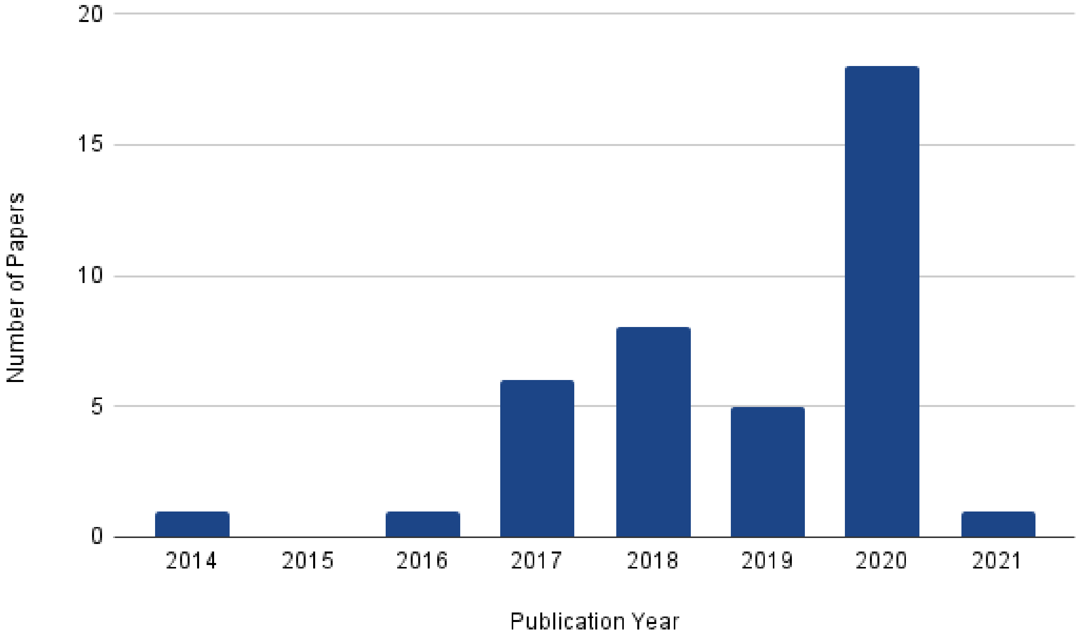

Fig. 3. Number of publications per year presenting a proposal on the use of blockchain in genomics. The bars show the number of papers for each year

4.2 Blockchain in Genomic Applications

This section reports the range of existing blockchain-based solutions in genomics, which answers RQ2.

4.2.1 Classification of the selected papers. Because the use of blockchain in genomics has attracted the attention of both academic and industrial communities, each with their own agenda on how this technology can be used, we classified the applications as shown in Figure 4 in two main categories: commercial and non-commercial applications. To clarify our classification process further, we distinguish between the two categories based on whether the blockchain is utilized for financial exchange in addition to data sharing. Commercial genomic marketplaces follow a business model and are generally aimed at facilitating the exchange of genomic data for financial benefits. In addition, these marketplaces are usually targeted at individual genomic data owners (individual users/customers or patients), and cryptocurrency or tokens are often used as incentives to promote data sharing. From our total number of selected papers (40), there are 27 papers with no commercial interests and 13 papers with a commercial motivation. Figure 5 shows the percentage of commercial vs. non-commercial papers on blockchain in genomic applications.

4.2.2 Commercial Genomic Marketplaces. The commercialization of DNA sequencing by direct-to-consumer (DTC) companies has attracted an increasing number of customers in recent years. This is due to the technological advances

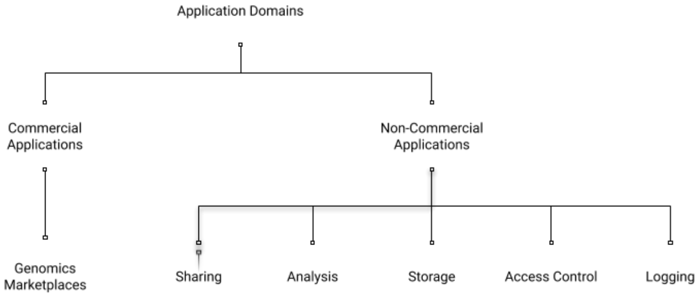

Fig. 4. Classification of genomic blockchain applications

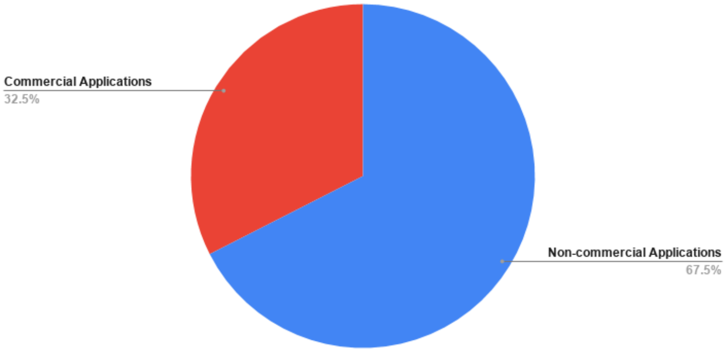

Fig. 5. Commercial vs. Non-Commercial papers.

that made it much cheaper and faster. One of the strategies to generate revenue for DTC companies such as 23andme is selling access to the collected DNA sequences to pharmaceutical companies. The fairness of this model raises questions in terms of the profit gained from buying this genomic data. Some argue that the profit should be passed onto the people, not the intermediaries [110]. As a result, there has been an increase in a new generation of companies that provides an open marketplace for genomic data sharing with the use of blockchain.

Blockchain-based genomic marketplaces aim to cut the need for intermediaries and give the users control of their data. Individuals receive different types of incentives for selling or renting their genomic data. The most common incentive used in these marketplaces is cryptocurrency. Table 3 provides an overview of the used incentives along with the employed blockchain platform and the offered services on current genomic marketplaces.

Genecoin [47] represents the first attempt at using blockchain in genomics when it was introduced in 2014. The company provides sequencing services through third-party labs. It then encrypts and stores the resulting DNA sequence in the bitcoin blockchain. The company does not provide any other motivation for genomic data collection rather than claiming to only gauge interest in this service in their website [47].

Table 3. Blockchain-based Commercial Genomic Marketplaces

Company	Blockchain Platform	Incentive	Services
Genecoin [47]	Bitcoin blockchain	N/A	backup encrypted DNA to Bitcoin blockchain.
Encrypgen [41]	GeneChain to facilitate cryptocurrency exchange and record data exchange	Cryptocurrency: DNA token	HIPAA-compliant storage of genomic data and exchange platform after de-identifying the data
Embleema [40]	Private Ethereum blockchain to register patient consent and manage data flow	Cryptocurrency: RWE token	Patient consent management, optimize clinical trial recruitment, secure HIPAA-compliant storage, analyze, and organize the data for researchers.
Nebula Genomics [35], [53], [54]	Exonum [43] permissioned blockchain for decentralized access control logs, data storage in third-party storage system	Nebula credits which can be used to purchase health-related information about the genome sequence	User controlled access to genomic data with blockchain log, secure storage and analysis in a blackbox environment.
LunaDNA [80]	Unspecified	Company shares and earnings when data are used	HIPAA-compliant storage, user ownership and control of data, contributions to research and medical research.
Shivom [130]	blockchain agnostic based on combination of blockchain implementations	Cryptocurrency: OMX tokens	Pay-per-use marketplace for genomic data, bioinformatics platform with analysis pipelines for researchers.
Zenome [95]	Ethereum-based blockchain	Cryptocurrency: ZNA token	Secure storage, selling access to genetic data, and buying genetic services.
Genomes.io [50]	Ethereum-based Blockchain	Cryptocurrency: GENE token	Secure personal DNA vault, and financial benefits from contributing anonymously to medical research.
DNAtix [37]	Ethereum-based blockchain (proof-of-concept)	Cryptocurrency: internal ERC-20 based token	Anonymous genetic vault service, anonymous testing and reports on genomic data, and connect anonymously with people with similar genetic traits.
Genesy [20]	Private blockchain based on HyperLedger Fabric	Payment in fiat and cryptocurrency through Stellar [116] and Stripe [119]	Sequencing services, selling access to genomic data, and a blockchain-based ecosystem for the sharing of genomic data.
GenoBank [127]	Ethereum-based blockchain with non-fungible tokens (NFT)	Cryptocurrency: ERC-20 token	Control over genomic data with DNA crypto wallet, and secure platform to process the data.
Longenesis [79]	Bitfury's Exonum blockchain	N/A	Patient consent platform targeted to medical institutions, HIPAA and GDPR compliance data storage,
LifeCode.Ai [63]	Quorum platform [28]	Cryptocurrency: ERC-20 token	Individual health data ownership, trading mechanism for data exchange, and secure decentralized storage and management of health data

EncrypGen [41] with its Gene-Chain allows genomic data sellers and buyers to exchange money in the form of a cryptocurrency (tokens) called DNA or more recently mDNA and eDNA (the ERC20 version of the token). Individuals willing to sell their data on the Gene-Chain can earn DNA token which later can be traded in cryptocurrency exchanges.

Embleema [40] offers a marketplace for cancer patients, and lets them control their data. The company aims to accelerate the oncology clinical research and at the same time incentivize users to participate with tokens (RWE tokens, which stands for Real-World Evidence). The company is in position of large number of patient records which is estimated to be over 50,000. In addition, it is using HIVE [113] (High-performance Integrated Virtual Environment), a parallel distributed computing environment, to accelerate drug discovery and FDA approval for new drugs.

Nebula Genomics [35] uses another business model, whereby people can upload their phenotypic data to the blockchain and earn tokens for doing so. They can then use these tokens, when they are enough, to purchase a whole-genome sequence from Veritas Technologies, which is a partner company of Nebula Genomics. There are also alternative ways: people can pay out of their pocket for the sequencing or a third-party, such as a pharmaceutical company or research center, can subsidize the cost of sequencing. Of course, the last option is only possible for particular health profiles which are attractive enough for these companies.

LunaDNA [80], plans to use blockchain as a marketplace for genomic data. Although the use of blockchain is still under development, they have already joined the Genetic Alliance, an advocacy group, with the goal to store this data in a cloud-based platform with security and privacy protections including access control and anonymization [30]. The company does not offer sequencing services but collects existing genomic data. In addition, it adopted a different model to incentivize users to share their data. Users receive company shares and therefore, they are part owners of the company and receive dividends once the aggregated data has value. LunaDNA believes that they are creating a community of part-owners, and in this community, the currency is the data.

Shivom [130] is a blockchain-based ecosystem with libraries and data pipelines that are specific for genomic data. The platform connects researchers with DNA data that are controlled by individuals. The data owners are first anonymized and researchers can then leverage this data to conduct their research through the provided pipelines. The company aims to first protect patient data and accelerate medical and pharmaceutical research.

Zenome [95] uses the Ethereum blockchain and its smart contracts to provide a marketplace for individuals to share or sell the right to access their genomic data to any interested parties such as researchers. The platform also allows users to store or buy computational resources from specific nodes in the network. These nodes are then rewarded with tokens called ZNA which stands for Zenome DNA tokens.

Genomes.io [50] is another genomics blockchain company that allows consumers to securely store and manage their DNA data from the moment it is sequenced to when it is stored on the blockchain. This prevents any attempt to tamper with the data and at the same time gives data owners the chance to monetize their data.

DNAtix [37] offers a blockchain solution that allows users to anonymously and securely store and trade genomics data. Patients' DNA sequences are submitted to the blockchain platform and then compressed using a proprietary algorithm. The proposed compression algorithm can reduce the size of genomic data and can address the data storage challenges associated with the growing amount of genomic data.

Genesy [20] aims to encourage data owners and organizations to collaborate by providing an ecosystem for managing the exchange and access to genomic data. Genesy provides sequencing services and the ability to sell access to generated data. The payment for these services is done through third-party APIs, namely Stellar [116] and Stripe [119], which allow both fiat and cryptocurrency transfers. Genesy utilizes a private blockchain based on hyperledger fabric [18] with the aim to grow beyond that and become a consortium blockchain managed by various organizations. The Genesy blockchain consists of multiple nodes that record the transactions as well as data. Sensitive data that includes the user's personal data are encrypted and stored within the blockchain, while other larger genomic data are stored off-chain on external databases and cloud storage with hash pointers on the blockchain.

Genobank [127] is exploring the use of non-fungible tokens (NFT) for portability and data tracing. The proposed method is to assign each unique human genome a unique NFT. The NFT allows for the full control of the user data while at the same time enables the data owners to authorize data consumers (researchers) to perform analysis on multiple environments.

Longenesis [79] is still under development and aims to provide a decentralized end-to-end marketplace for health data including blood test results, medical history and genetic profile. Users will be able to use the Longenesis's platform to store and consent to participating in a specific medical study. The users could withdraw their consent at anytime. In addition, using smart contracts, the medical providers can offer to extend, modify or amend an agreement which can be accepted or rejected by the users.

LifeCODE.ai [63] is a blockchain-based platform with a focus on storing and managing genomic data. The decentralized application (DApp) created by LifeCODE.ai facilitates trading of data through Ethereum's ERC-20 protocol that is implemented in the Quorum blockchain. The tokens are used to pay for access to patient data. To protect the privacy of the data, all health data stored in the blockchain network are encrypted. In addition, the data are owned by the individuals that submit them and all data movements are traceable.

4.2.3 Non-Commercial Applications. The selected studies on non-commercial applications of blockchain mainly focus on providing genomic data sharing for the advancement of research. Table 4 provides the list of scientific works we identified that applies a form of blockchain in relation to genomics. The selected studies fall into one or multiple of the following subcategories: data sharing, analysis, secure storage, access control, and logging. While each paper is categorized according to the main topic of research, overlaps occur. For example, one study focused on genomic data sharing for the purpose of performing analysis tasks on that data, and another paper focused on blockchain-based storage for the purpose of sharing the stored genomic data. In our analysis, we account for this overlap (as seen in Table 4).

A. Data Sharing

The majority of the papers we identified focused on using blockchain to support/build systems for multi-organizational or global sharing of genomic data. A noteworthy paper is the cancer gene trust (CGT) [52], which demonstrates the benefits of blockchain in sharing genomic data, for the purpose of advancing cancer research. In addition, the authors launched a cohort study with a real patient dataset to illustrate the effectiveness of the CGT framework in terms of secure, efficient, cost-effective, open, and distributed sharing of genomic data. A similar approach is presented in [60], but with additional mechanisms to distribute the whole-genome data.

B. Data Processing/Analysis

We also identified a set of papers that explored the use of blockchain in facilitating genomic data processing or analysis. Zhang et al [135] proposed an approach to perform a Genome-wide association study (GWAS) with a focus on privacy. The authors proposed performing GWAS by using a privacy-preserving sharing protocol (PPS) that enables genomic data sharing through the use of a gene fragmentation framework. The large genomic files are split into multiple fragments which are then distributed in a decentralized blockchain network to multiple service providers for storage, sharing, and analysis. This eliminates the possibility of one provider having the complete data, therefore, solving the issues related to centralization and privacy protection. Coinami [60] provided an alternative approach by incentivizing participants to perform HTS read mapping. The participants are given tokens as a reward. This replaces the traditional proof-of-work with HTS read mapping to validate blocks in the blockchain. Other approaches proposed combining genomic predictive modeling with blockchain to achieve a distributed model training. In [69, 70], predictive models were trained in multiple organizations with blockchain coordinating the process in a decentralized way.

Verification of the computation and analysis tasks performed by third parties is essential. In a blockchain setting, it is important to check the validity of the computation done by untrusted nodes in the network. In addition, the verification process should optimally be done with minimal computation resources. Zhang et al. [135] use a blockchain consensus protocol to validate the results of analysis or computation. Each job is assigned to multiple nodes and the outcome is

Table 4. Blockchain in genomics applications.

Paper	**Application Domain**				
	Data Sharing	Data Analysis	Secure Storage	Access Control	Logging and Auditing
[135]	✓	✓			
[98]	✓				
[52]	✓				
[57]	✓		✓		
[84]				✓	
[82]					✓
[78]	✓		✓		
[136]	✓	✓			
[123]	✓				
[60]	✓	✓			
[56]	✓	✓	✓		
[31]	✓		✓	✓	
[55]					✓
[112]	✓				
[62]	✓	✓			
[86]	✓		✓		
[94]	✓				
[71]	✓	✓			
[69]	✓	✓			
[70]	✓	✓			
[97]					✓
[73]		✓			
[101]					✓
[141]			✓		
[42]	✓	✓			
[16]				✓	
[99]	✓		✓		

compared. Then the result of the majority is considered correct. In [60] the computation results are checked by certified authority nodes in the blockchain network. The validity is ensured by randomly inserting pre-calculated decoy data.

C. Secure storage

Among the selected papers, we found a group of papers that focused on using blockchain as a way to store genomic data securely. [57] utilized blockchain to store and query pharmacogenomics data. The authors illustrated the feasibility and efficiency of storing and accessing this data using the Ethereum blockchain and smart contracts. Each data record is inserted into the smart contract and assigned a unique ID to be used as mapping key. An index-based, multi-mapping approach is used to efficiently query the genomic data. The pharmacogenomics data used in this study is rather small in size compared to other common types of genomic data types. In [56], the authors explored storing larger data files, specifically Sequence Alignment Map (SAM) files, which can be in the order of 10s of Gigabytes in size. This was achieved with a novel data structure that was built with the addition of data compression techniques and a private blockchain network.

D. Access Control

We found a small number of papers with a focus on using blockchain as means to provide and revoke access to genomic data in the form of consent management. Dwarna [84] is a web portal that harnesses blockchain for dynamic consent. The portal connects participants and researchers in a research partnership. The project incorporates GDPR and gives ownership of the data to the participants. The proposed architecture uses blockchain to record participants' consent. Storing consent in blockchain allows the participants to be the owner of the data i.e. third parties can only access the data when the owner of the data allows them to. In [16], the authors consider consent for sharing individual genomic data as an instance of the Multi-Stakeholder Consent Management (MSCM) problem. This is due to the fact that each individual genome can reveal information not only about the owner of that genomic data but also about relatives. Therefore, to protect the privacy of the relatives of an individual, their consent must be taken into account. The authors in [16], propose the use of blockchain to solve this consensus problem and obtain consent from multiple stakeholders.

E. Logging and Auditing

Another set of papers focused on the value of blockchain in building a global logging system. The 2018 IDASH competition [72] and specifically, Task1 of the competition explored the use of blockchain as a global logging system. Such a system can be used to provide an access log that records users' access to any data within any of the genomic data repositories in the system. A decentralized cross-site logging system has many advantages over traditional centralized internal logs that are currently common in practice. Most importantly, it eliminates the problem of a single point of failure and malicious changes to the logs. There were several participants [55, 82, 97, 101] in the competition and the submissions were evaluated based on specified criteria which include accuracy and speed. This is because the competition not only looked at the feasibility of blockchain as a cross-site logging system but also evaluated its performance and efficiency. The winner of this competition [55] showed that it is indeed feasible to utilize blockchain in building a cross-site genomic data access log. The performance of that solution is promising, and it is reasonable to assume that with additional improvements, such a system can be adopted for practical use.

4.3 Motivations for the use of blockchain

To address RQ3, we first identified the key blockchain features that are most desirable in genomic applications. These key features are incentives, decentralization, control of data, immutability, smart contracts, reliability, availability, transparency, and traceability. The motivation is different for each genomic application and it depends on the outlined requirements that are listed in each paper. Figure 6 shows the frequency of each key blockchain feature that motivated its use in the selected papers. Most papers list multiple benefits of using blockchain. The most highlighted feature is that it is an immutable and tamper-proof way to store data. In addition, decentralization and control of the data are highly mentioned benefits. The rest of this section provides a summary of the motivations to apply blockchain in the selected studies as described by the authors.

A. Incentive (Cryptocurrency). An important mentioned benefit of using blockchain is the ability to build an incentive structure for sharing genomic data. This is especially relevant for genomic marketplaces where the objective is to create a fair ecosystem for the exchange of private data. The fairness is defined in terms of financial gain from data sharing (by data owners), and the aim is to allow the exchange of data for scientific research or other purposes without losing full control. Blockchain provides the required incentive structure for distributing genomic data in exchange for cryptocurrency or tokens. Additionally, an incentive scheme can be used to reward nodes in that network after completing a certain task such as sequence (HTS) read mapping in [60]. Any individual or organization is able to freely join and perform analysis tasks to gain tokens that could be redeemed for monetary value.

B. Decentralization. The decentralized nature of blockchain networks was listed as an important feature in several papers. The consensus mechanism contributes greatly to the way blockchain is decentralized. It introduces a way for nodes in the network to reach an agreement without a central trusted authority. Decentralization provides several benefits depending on the use-case. In [52] decentralized open access is achieved by using blockchain combined with IPFS. Timely distribution of medical resources plays a significant role in the research and development of medical treatment especially in the event of a disease outbreak such as COVID19. However, the technical limitations in applying reliable decentralized technologies is one of the barriers in achieving this, which have kept such valuable data in silos and behind centralized servers. The authors in [52] claim that blockchain fills that gap and provides a reliable decentralized data sharing system. Decentralization is also useful in other cases such as coordination of analysis tasks that are performed at multiple nodes/locations. As shown in [69–71, 73], blockchain can replace the need for a central server to intermediate the process of applying machine learning and combining the global model. This prevents the single point of failure/control when a third-party is used to coordinate, and the potential for this third-party to breach the privacy of data by examining the aggregated statistics. A similar approach is shown in [135] where the motivation for using blockchain is to coordinate the process of performing GWAS studies and guarantee the authenticity and confirmation of all activities (transactions) within the decentralized network.

C. Control of Data. Control of genomic data should ideally be given to the owner of the data (the patient) or a trusted third-party acting on behalf of the owner such as doctors. Necessary consent and access management mechanisms in the current centralized systems requires more time and effort. A set of papers list the ability to control of data as the main motivation for using blockchain. This is especially evident in genomic marketplaces that claim to allow individuals to control who has access to their data and for what purpose. As mentioned previously, there are also proposals (e.g. [85]) in which the patient consent is stored on the blockchain to empower patients and enforce their control over their own data.

D. Immutability. Immutable and tamper-proof data storage is the most desired property of blockchains in the selected papers. The immutability property in blockchain prevents the loss and alteration of data records which is essential in most genomic applications. The tamper-proof data structure of the blockchain, which relies on cryptographic hash pointers, prevents both accidental and intentional data tampering. Any changes to the confirmed blocks would make the blockchain inconsistent and can be discovered by any node in the network. This ensures a reliable and consistent shared ledger among untrusted or semi-trusted parties in the networks. Depending on the application requirements, on-chain storage can be leveraged to store data that needs to persist such as recording consent [85] and providing an audit trail [55, 82, 97, 101]. However, data privacy must be carefully considered since the immutability property applies to on-chain data that is shared across all nodes and can be openly viewed if not encrypted.

E. Smart contracts. Smart contracts support various functions such as token generation and distribution, access control, and policy enforcement. Smart contracts are not part of all blockchain platforms, and not all applications require them. However, a percentage of papers have included smart contracts as a major part of their proposals. Commercial genomic data exchange platforms utilize smart contracts to generate the tokens that are used to incentivize individuals to share their genomic data. Both Encrypgen [41] and DNAtix[37] use smart contracts to generate ERC-20 tokens that are used for payments. Smart contracts can also support the trading of genomic data for these generated tokens. In addition, enforcing an access policy by utilizing smart contracts is another motivation for employing blockchain. For example, [20] relies on smart contracts to allow access to the data and transfer the payment to the data owners.

F. Reliability, availability, transparency and traceability. A small percentage of papers specifically listed these blockchain properties as the main motivation for using blockchain. Reliability and availability are essential in certain applications such as online model learning in [70], which require data to be highly available to all nodes in the network. The importance of the transparency and traceability are mostly highlighted in applications where the data owners are informed of how the data is accessed and by whom. These blockchain properties can be exploited to give data owners control and in turn gain their trust. For instance, [85] emphasized the importance of transparency and argued that patients are more willing to contribute their genomic data for research purposes when they are informed about the use of their data.

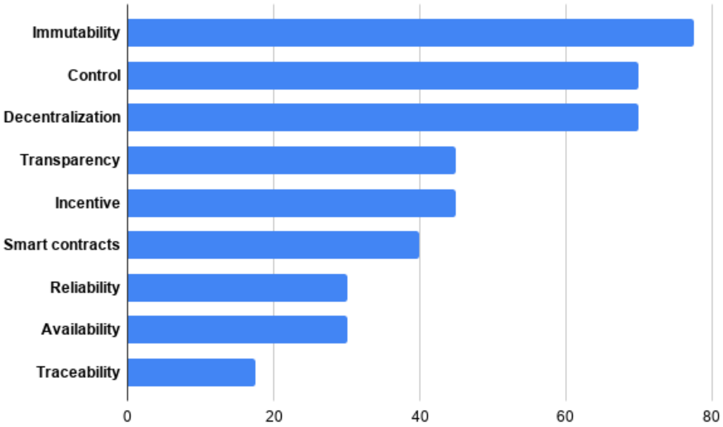

Fig. 6. Frequency of motivations for the use of blockchain in genomics.

4.4 Current Blockchain-based Methods and Approaches in Genomics

There are various platforms [74], storage systems [13], and privacy-preserving techniques [15] tailored to blockchain-based genomic data solutions. These technologies can be combined in different ways to deliver an application. In this section, we look at different approaches used in genomic blockchains in order to completely answer RQ4. First we present a general architecture covers most of the existing genomic blockchain systems. We then use this architecture to guide our presentation of the discussed work.

4.4.1 General System Architecture. Most previous work such as [133, 137] have discussed the architecture of blockchain in a general way, while others discussed the architecture in a specific application such as IoT [75]. In this section, we present an application-specific architecture tailored to blockchain-based genomic data applications.

Blockchain technology has been used for a variety of genomic applications as we have listed in the previous sections. Each of the selected studies has its own system architecture for the specific application at hand. However, most of the proposed solutions have architectural similarities that paves the way to generalization. In Figure 7, we present a generalized architecture for systems that use blockchain for genomic applications. The design for this architecture was aimed to summarize and cover a wide range of applications in genomics. Our proposed architecture consists of 6 layers:

data collection layer, data storage layer, network layer, consensus layer, application layer, and presentation layer. The layers in this architecture are comparable to the one/s in the existing blockchain literature but with some modifications to effectively illustrate the architectural components exploited in genomic applications. [94] proposed a similar system architecture that consists of three layers: data gathering, storage, and application layer. We extended this architecture to include all layers within the blockchain.

In the first layer of the architecture, we assume that each node is responsible for collecting and sorting the genomic data which can come in different formats such as BAM or FASTQ. These nodes in the systems can represent an individual, researchers, or organizations that want to share genomic data. Individuals can submit their own genomic data after being sequenced, which is the aim of most genomic marketplaces. Researchers and organizations can also obtain consent from patients to release their genomic data, or an anonymized version of it, to other researchers in the context of a particular disease. This is the case in [52], where the de-identified genomic and clinical data are collected from cancer patients after consent is given. There are also other data types such as gene-drug interaction in [57], and patient consent data [85].

After the collection, the data is sent to the next layer for storage. In the storage layer, the data can be stored in different ways depending on the requirements, which is discussed in section 4.4.3. The data is then broadcasted to the network using a specified network protocol. According to our analysis, the majority of the papers use a P2P network instead of the traditional client-server model. In the consensus layer, the nodes in the network come to an agreement on the state of the blockchain using a consensus protocol such as Proof-of-Work (PoW). The application layer is where smart contracts are written and deployed to facilitate various application functions which serve as the backend of the application. The presentation layer is responsible for interacting with smart contracts and blockchain in general.

4.4.2 Blockchain Platforms. A critical step in designing a genomic blockchain system is the selection of a suitable blockchain platform that would deliver the required functionality for the intended application. In this section, we present our analysis of the blockchain platforms used in the selected studies. Our analysis is only based on solutions that included a prototype, a proof of concept, or an implementation to show the feasibility of blockchain in genomic applications. There are also some genomic blockchain studies that do not explicitly specify the blockchain platform or implementation, and others that do not reveal their underlying platforms especially in commercial applications. These are not covered in our analysis.

Despite the fact that there are multiple types of distributed ledgers such as Directed Acyclic Graph (DAG) and Hashgraph, blockchain is the only type discussed in the genomic literature to our knowledge. In addition, the majority of the papers proposed the use of either private or permissioned blockchains. Privacy, scalability, and cost are among the most cited reasons for this. The use of private blockchains lowers the risk of information leakage since data is only shared with a set of known semi-trusted individuals or institutions. In addition, private blockchains are more scalable and often use consensus mechanisms that do not require cryptocurrencies and transaction fees. Aside from custom-made blockchain implementations that are designed to fit specific application requirements, the two main platforms used in genomic blockchain solutions are the following:

MultiChain [45] is a blockchain platform to build and deploy private or permissioned blockchain networks. MultiChain focuses on providing features such as mining diversity, round-robin-based consensus, and data streams that allow on-chain data storage in a secure and efficient way. However, Multichain does not support smart contracts as do some other blockchain platforms. Multichain was used to build GeneChain [41]. It was used in the IDASH competition [72]

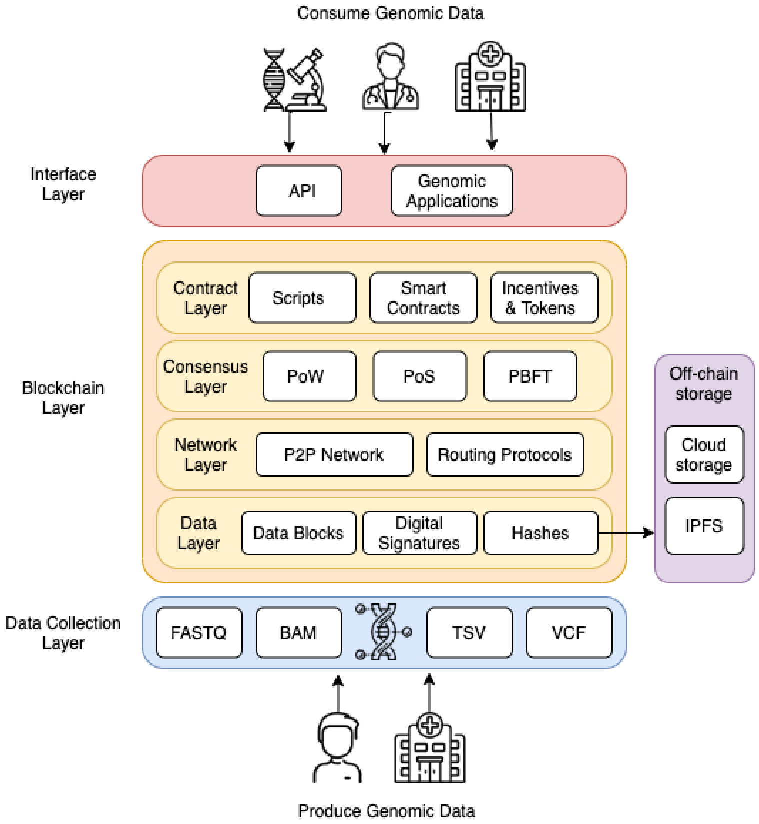

Fig. 7. General architecture for blockchain-based genomic data sharing systems

and thus all presented solutions [55, 82, 97, 101] are based on MultiChain. The use of Multichain in these papers allowed efficient on-chain data storage of access logs. Additionally, MultiChain was used to implement ExplorerChain [69, 70] for the purpose of distributing the online machine learning models to the nodes in the permissioned blockchain network.

Ethereum [27] is a blockchain platform that facilitates building smart-contracts and decentralized applications (dapps) that run on the blockchain network. Ethereum focuses on adaptability and flexibility, and, to achieve this, it supports Turing-complete programming language to build smart contracts easily. Although the Ethereum main blockchain network is public, in most of the selected papers, a private Ethereum blockchain is used. In multiple genomic marketplaces such as [35, 37, 95, 127] an Ethereum-based blockchain was used. In these use-cases, Ethereum smart contracts are used to facilitate access to genomic data files and the distribution of cryptocurrency. The previously mentioned CGT framework [52] is another example solution that relies on Ethereum smart contracts for the distribution of genomic data.

4.4.3 Storage Techniques. There are two types of data storage that are compatible with blockchains, namely: on-chain and off-chain. The use of one or a combination of the two depends on the design requirements and consideration. This section describes the current blockchain storage approaches used in the literature to facilitate sharing and management of genomic data.

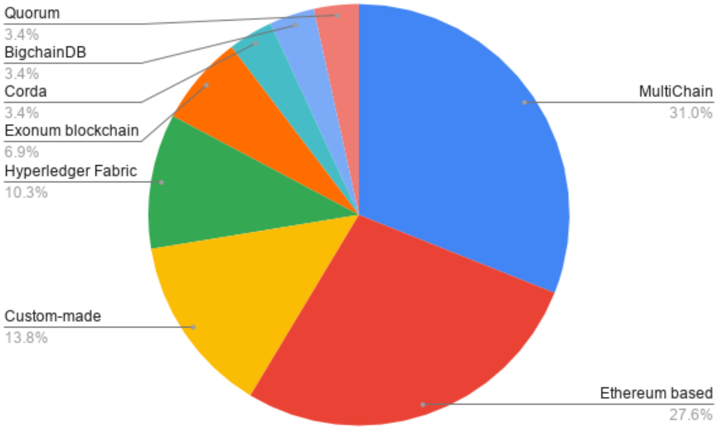

Fig. 8. Blockchain platforms used in genomic data applications

On-Chain Storage. Storing data on-chain is achieved by simply adding the data (in binary format) to the transaction which effectively makes it part of the chain itself. This will eventually make the data immutable and highly available as the transaction will be distributed to all nodes in the network. However, some blockchains, especially public blockchains have a strict limit on the size of each transaction making it difficult to store large data files. This is due to the fact that each full-node needs to have enough resources to store the ever-increasing amount of data being generated. On-chain data are also publicly accessible to all nodes in the network and therefore privacy of the stored data must be considered.

On-chain storage is most suitable for small data types that require immutable and tamper-proof storage. Data types that are commonly stored on-chain are meta-data and small genomic data. Small data types such as audit trail and observations of gene-drug interactions can be effectively stored on-chain as shown in [55, 57, 82, 97, 101]. However, the majority of the papers only use on-chain storage for metadata [52, 69, 70, 73, 85, 98].

On the other hand, whole-genome data which are stored in files such as BAM or VCF are large in size and difficult to store on-chain. There are attempts to store this data on-chain, such as genecoin [47], and SAMchain [56]. Genecoin [47] sends DNA kits to customers, and after sequencing the DNA sample by a third-party sequencing facility, the extracted full-genome data is encrypted and stored in the bitcoin blockchain. While this approach is possible, it is not feasible for public blockchains. Speed and scalability are severely affected by the need for each node to replicate these large data files. On the other hand, SAMchain [56] uses a private blockchain (Mutichain) to store and share sequence alignment maps on-chain using nested database indexing and compression techniques. Indeed, one of the the main motivations of this paper is to prove that efficient storage and analysis are possible with a private blockchain.

Off-Chain Storage. The practical limitations of on-Chain storage can be overcome by utilizing off-chain storage. In general, most off-chain storage techniques involve hashing (a piece of) data, which results in a small string that can be efficiently stored in the blockchain transactions or in a smart contract [13]. The actual data is then stored in local centralized storage system or may be replicated between multiple nodes. Smart contracts and distributed hash tables (DHT) are two of the most common approaches for off-chain storage. Smart contracts can specify what the data is, who has access to it, and where it is stored. On the other hand, a DHT is a network of storage nodes with a centralized index. The index stores the information that points to where a specific piece of data is stored. The storage nodes can either store the data entirely or a piece of the distributed data. One of the most popular DHT solution is the InterPlanetary File System (IPFS) [61].

An example is the CGT framework which enables sharing of large genomic data in a distributed environment through the use of off-chain storage, namely, IPFS [61]. Data is stored on the IPFS servers and only a strong hash (SHA-256) is added to the blockchain. The hash uniquely defines the entire state of all data submitted from the steward at that exact point in time. Similar hashing of raw data is used in [136] to validate that off-chain data has not been tampered with. Another notable off-chain storage technique, employed by Genesy [20] and CrypDist [98], is to use cloud storage and linking the data to the data owners through cryptographic pointers stored in the blockchain. A transaction is created with every file containing a hash pointer to the data on the cloud. What is stored off-chain in this scenario is large BAM files, while access metadata and patient metadata, such as phenotypic and environmental data are stored on-chain.

Compression. Data compression is an alternative approach to overcome the data storage limitations in blockchains. Compression reduces the dependence on large storage and the time/cost of transmitting large genomic data. [78] proposed a lossless compression algorithm called Blockchain Applied FASTQ and FASTA Lossless Compression (BAQALC) that provides efficient storage and transmission of next generation DNA sequence data on a blockchain network. In addition to early mentioned PetaGene [102], which claims to achieve 60% and 90% savings for BAM and FASTQ formats respectively, there is the MPEG-G [91] initiative. While blockchain relevant storage systems such as IPFS may have their compression methods, it is intuitive to think that more and more genomic data specific compression techniques, as in [78], will be employed in blockchain storage systems themselves.

4.4.4 Security and Privacy Protection. In genomic applications, the privacy requirements vary depending on the data type. Somatic variants are generally considered to be non-private and do not require any privacy protection. On the other hand, germ-line variants are private and therefore, privacy protection is essential and it is enforced by existing regulations such as HIPAA and GDPR. In addition, genomic and medical information (extracted from EHR) are often combined for genotype-phenotype analysis. This creates another point of privacy concern because when multiple data points are combined, there is a risk of correlation attacks and patient identity can be revealed by combining multiple identifying data points [93]. With these privacy concerns, it is necessary to look closely at how this is currently achieved in genomic blockchains. In this section, we look at the security and privacy aspects of genomic blockchains.

The majority of the papers rely on the security of the cryptographic protocols employed in the basic blockchain implementation which include Hash pointers, Merkle trees, digital signature, a public key infrastructure (PKI), and a consensus protocol [133]. As previously discussed in Section 3.2.4, the combination of these techniques provides a robust decentralized system that can withstand malicious tampering.

Privacy through data anonymization is one of the approaches used in genomic blockchain literature, both CrypDist [98] and CGT [52] achieve privacy by sharing only somatic variants and removing the personally identifiable patient data and private germ-line variants. While anonymization provides a certain level of privacy, it does not guarantee protection against future re-identification attacks [106].

Another approach to privacy is the use of private blockchains. Gursoy, et al [56] uses a private blockchain that requires permission to access the data within the blockchain. With controlled access, there are limited number of security issues. According to the authors, it is also possible to store homomorphically encrypted data in SAMchain, which allows privacy-preserving computation on the data. However, the efficiency of this was not addressed.

[69–71] address the privacy concerns of sharing patient data by distributing machine learning models to multiple institutions rather than sharing the data itself. The authors use blockchain as a way to coordinate the process of distributing the model instead of a central server that could potentially breach the confidentiality of the data. However, the authors point out that risks of re-identification still exist and further privacy-preserving techniques such as differential privacy are required for optimal privacy protection.

Genie [136] presents a blockchain-based solution to AI model training with the added security of a trusted execution environment namely, Intel Software Guarded eXtensions (SGX). The secure enclave is used to train the models and therefore privacy is preserved by protecting the raw data while still allowing the sharing of insights from it. These security and privacy protection techniques are combined with the transparency, control, and verifiability of blockchain in the proposed solution.

Zhang et al [135] provided a blockchain approach to perform GWAS studies with privacy protection. The analysis is done through third parties which provide the computing resources required to perform the analysis. Privacy preservation is achieved through a novel gene fragmentation framework. In the proposed framework, the gene sequence of one individual is fragmented into *n* pieces and distributed to analysis nodes, which run a specified analysis on the given part of the data. The fragmentation lowers the probability of re-identification in each analysis node and makes sure each fragment of the data is unidentifiable.

[54] uses multiple cryptographic techniques that are added to the data-sharing system. Homomorphic encryption is used to encrypt the data and process it. Differential privacy is also used to add another layer of privacy and prevent re-identification of individuals. This is done by adding noise to obfuscate the query results.

4.5 Open Issues and Challenges

The following is a list of current challenges and limitations in applying blockchain in genomics which answers the predefined RQ5.

4.5.1 Adoption barriers. Public adoption of blockchain in genomics and healthcare, in general, requires ease of use and software stability. Blockchain platforms are continuously changing and require expert knowledge for adoption. This has been noted as one of the major challenges in implementing solutions in most papers. Instability and lack of user-friendly interfaces are major barriers to public adoption and therefore limit its use to tech-savvy individuals.

4.5.2 Interoperability. For the adoption of blockchain in genomics, organizations need to integrate and connect blockchain platforms with existing non-blockchain platforms. This problem is aggravated as there is a wide range of blockchain implementations that are not necessarily compatible with each other. Interoperability between blockchains reduces the dependence on a single blockchain platform. Most of the presented solutions in genomics rely on a specific blockchain platform and its features. A multi-blockchain approach, which doesn't rely on a specific blockchain platform, would provide better scalability and remove the security risks associated with the used platform. However, this approach is currently complicated and involve complex cross-chain communication. Research in the area of blockchain interoperability is growing and multi-blockchain approaches might be feasible in the future [12].

4.5.3 Smart Contract Security. Blockchains that support smart contracts enable building rich applications that are not limited to financial transactions. However, the increased functionality provided by smart contracts exposes the system to more possible attacks such as the DAO attack [88] in 2016. The number of discovered smart contract vulnerabilities is increasing [23], and they can be costly, either in terms of financial loss or data privacy loss. Smart contracts were used in a number of papers to achieve various important functionality, such as granting access to private data. One of the challenges in deploying such smart contracts to handle actual patient data is the security risk associated with them.

Following best practices and performing security audits might reduce this risk, and research in smart contract security is still ongoing [23].

4.5.4 Data Privacy. Even though privacy in blockchain has been studied extensively in the literature, privacy issues with blockchain in genomic applications have not been fully addressed. There is a need to examine some areas of privacy especially with the anonymity of users and re-identification through correlation attacks. While re-identification is sometimes required in research settings, it is essential to prevent disclosure of patient data for any other purpose. For instance, re-identification is required when additional materials are needed for further research into the case. However, re-identification to reveal patients' private data should be prevented. The privacy challenges in genomic blockchains include the following:

(1) Identity and transaction privacy: maintaining the user's private identity and not relating it to the transaction. Correlation attacks, in which true identity could be revealed, is a privacy challenge in using public blockchains that requires further research. Ideally, identifying a user based on specific interactions with organizations should not be possible. On-going research in using zero-knowledge proofs (ZKPs) [100] have demonstrated the possibility of achieving this in financial applications.
(2) Re-identification risks: in the case where blockchain serves as open access to genomic data for research purposes, the process of anonymizing and obtaining consent for genomic data is time-consuming and requires an honest third party as has been observed by [52]. This is difficult to scale when the number of patients is large. Moreover, risks of re-identification associated with open data sharing are still present even after the full de-identification process as has been shown in [106]. While this process might follow the best practice in anonymizing the data, there is still a risk that more advanced re-identification attacks might emerge in the future. Therefore, it is an open question whether other privacy-preserving mechanisms can be applied to ensure privacy against future attacks.

4.5.5 Reproducibility and validity. Replicating the study to determine its validity is essential in genomic and research in general. Researchers or auditors would want to find exactly the same data without change after a number of years to replicate the study. In cases where blockchain facilitates data sharing for scientific studies, it is important to address this since data in a decentralized network reside in multiple storage locations. This problem manifests especially in solutions that use off-chain storage. The data should follow the FAIR principles, however, there seems to be a lack of focus on this aspect in genomic blockchains.

In addition, data redundancy might occur where patients exist in multiple organizations with different assigned IDs. A challenge with decentralized sharing/analysis of genomic data is the possibility of publishing duplicate data, where the same patient data is shared but with different anonymous identities. This is a problem especially if this data is intended for research purposes and it can affect the validity of the study. The Cancer Gene Trust [52] highlighted this limitation and tried to eliminate this problem with a rule-based scoring system and 2 reviewers. There are existing techniques to identify duplicate records in different databases, such as [76] which performs privacy-preserving record linkage on several databases using secure multiparty computation. However, further research is still needed to address this problem in a decentralized blockchain network.

4.5.6 Verifiability. One of the issues associated with distributing processing tasks to untrusted parties is verifying the accuracy of the results. One possible solution is outsourcing the same analysis task to multiple analysis nodes and then the results can be compared to ensure correctness. However, the cost of this approach can be high, especially if the same task is distributed to a large number of analysis nodes. Therefore, a practical and scalable verification method to ensure that outsourced computation/analysis is indeed correctly computed by untrusted nodes in a blockchain network is still an open issue that requires further examination. There are significant advancements in the field of verifiable computation [132] which can be explored in a blockchain setting. Cryptographic techniques such as homomorphic encryption [49] and zero-knowledge proofs [5] can be used to maintain verifiable results. These techniques might prove to be effective in overcoming the limitations in existing work in genomic blockchains.

4.5.7 Key management. It is challenging to ensure that the patients (data owners) are able to manage securely their keys and identity, especially with data related to individual health. Once patients have full control over their data, education mechanisms must be put in place for the patients, in order to provide them with valuable insights regarding best data management practices. Moreover, proper key management schemes need to be put in place along with mechanisms for "break glass" access to genomic/healthcare data in emergency settings.

4.5.8 Ethical Challenges. The rise of genomic marketplaces raises some ethical concerns as discussed by Ahmed et al [3] and Defrancesco et al [33]. These authors argue that informed consent is questionable when a monetary incentive is involved and it can lead to mindless data sharing. It is yet to be known if these financial incentives would actually work in attracting more people to share their private data for research purposes, but perhaps alternative non-monetary incentives should be explored. For Instance, Mofokeng et al [90] showed how digital collectibles can be used to incentivize citizens to participate in wildlife conservation. The authors, in collaboration with CryptoKitties, have created a non-fungible token (NFT) and a turtle-inspired CryptoKitty. Then, they put it for sale on the blockchain and raised 25,000 for the conservation of wildlife. The buyer holds an immutable and unique digital asset which marks their contribution to wildlife. Therefore, further research into non-monetary incentives that encourage participation in genomic research for the purpose of advancing medical research might prove to be effective.

5 DISCUSSION

In this section, we discuss the major findings of this review, the challenges, limitations, and propose several future research directions in applying blockchain in genomics.

5.1 Principal Findings

Our results suggest that there is an increasing interest in applying blockchain in genomics since the number of publications increases rapidly each year since the inception of blockchain and the developments in genomics. The use of blockchain technology in genomic applications is becoming common in both commercial and non-commercial settings. Commercial applications focus on the need for user-control (i.e. data ownership) and at the same time enable to profit from its use and sharing. Rewarding a form of cryptocurrency to the data owners is commonly used as an incentive mechanism to attract more contributions from the users. Non-commercial applications provide solutions for sharing, processing/analysis, secure storage, access control, and access logging of genomic data. These solutions aim to facilitate easy, efficient, multi-organizational genomic data sharing for the purpose of advancing the genomic research. The main motivations for using blockchain that were highlighted in the papers are *immutability*, *decentralization*, and *access/usage control.*

Our findings indicate that private or permissioned blockchains are the most common blockchain types, and Multichain and Ethereum are the most common platforms used in genomics. The storage techniques used vary depending on the requirements. *On-chain storage* is mainly used for small data types or hashes/pointers of the data that are stored off-chain. *Off-chain storage* is often used for large data files or for data that requires strict access control. In these cases, either cloud storage or other decentralized file systems such as IPFS is used. Further investigation of *data compression* techniques suitable for decentralized storage, sharing and processing of genomic data is needed.

Since genomic data is long-lived (i.e. valid for a very long time), data security and privacy are among the principal concerns. Most of the existing papers utilize existing protection mechanims that are part of the blockchain itself such as consensus mechanisms and digital signatures for the purposes of integrity, confidentiality and availability. Privacy is protected by either anonymization or the use of private blockchains with controlled access. There are, however, solutions that proposed adding cryptographic techniques to further protect the privacy of the data. More recent efforts aim to perform *distributed genomic analysis* through the use of smart contracts which we believe is a promising line of research due to better privacy guarantees and data retention controls.

5.2 Review Limitations

While this review aims to provide a comprehensive overview of the current state of the art, limitations still exist. The main focus of this review is to cover blockchain applications in genomics, therefore, we do not cover other healthcare-related studies in which blockchain is proposed as a solution. Our search strategy was aimed to capture only papers that specifically emphasize genomic data applications. We also focused on the most popular commercial genomic platforms, and therefore we only reviewed the first 200 results returned from the initial Google search. The documentation in white papers is somewhat limited and could change as the technology matures. We observed that details in some white papers change over time. We tried to overcome this by including the details of the latest versions of these white papers. Additionally, we have not run any testing on the proposed solutions to verify the claims stated in the papers.

5.3 Future Directions

To achieve large-scale deployments and adoption of blockchain in genomics, we point out that further exploration is needed for this technology to mature. There are multiple challenges and interesting problems in blockchain research that need to be addressed. In this section, we highlight some possible future research directions that we observed after conducting the review.

Based on our results, the use of cryptographic protocols for privacy-preserving analysis and computation is limited in current blockchain-based approaches in genomics. Cryptographic protocols such as Multi-Party Computation (MPC) can enhance privacy and when combined with blockchain and smart contracts, it can help address the security, trust, and verifiability issues in distributed analytics [138]. Current research in this field has shown the feasibility of performing distributed privacy-preserving analysis on blockchain [51, 139]. Researchers can further investigate and develop new approaches for genomic applications.

Furthermore, we observed that the solutions are all based on a limited number of blockchain platforms. While these platforms are popular and have proved to be effective, perhaps experimenting with other emerging blockchain platforms might show some interesting results. Another point to consider is that all papers included in this review propose the use of blockchain and not DLT. The use of DLT has been proposed for other applications such as IoT, but it seems to be missing in the genomic literature. Further research into the use of DLT is required to assess the feasibility of this technology in genomics.

Attribute-based encryption (ABE) is yet another powerful technique that can be leveraged to enhance the security and privacy in blockchain solutions. With ABE users with certain attributes defined in their secret keys can decrypt the encrypted data with matching attributes. [121] proposed the use of ABE in verifying the authenticity of electronic health records (EHR). However, combining ABE with blockchain has not been utilized in current genomic applications and future research can explore this largely unexplored cryptographic method in blockchain literature. For example, ABE can be used to grant specific attributes to certain nodes or users in the blockchain network which in turn allows them to access or perform specific tasks on genomic data.

Off-chain computation is another opportunity for future research. Recent development in moving smart contract computations off-chain is promising and would open the door for more applications to be built on the blockchain. There is incentive-based verification of off-chain computations such as Arbitrum [65], and others based on cryptographic methods such as zero-knowledge proofs [39]. We believe there is an opportunity for researchers in genomic blockchains to experiment with these methods and perform various genomic data analyses such as GWAS studies and determine their efficiency and privacy.

Genomic big-data are known to be difficult to move, and an emerging approach is to move the analysis pipelines to the data rather than moving the data itself. With this setup, organizations control the use of the data that resides in their own data repositories while sharing valuable insights from this data to researchers at the same time. Recent work, such as [69, 70], have shown the validity of this approach. However, this is limited to model training. Further research into the applicability of using blockchain for distributed analytics might also prove to be effective.

We also see a lack of focus on the aspect of trust and the emerging technologies that can be used to build it such as Decentralized Identifiers (DIDs) [104] and Verifiable Credentials (VCs) [115]. Utilizing blockchain to build a decentralized trust infrastructure, which can help in addressing other challenges such as data integrity and privacy. Following the trust-over-ip principles [32] in the genomic applications can provide a layer of trust when sharing data for research purposes, for instance, by sharing data only to those with credentials matching specific criteria.

The use of blockchain in genomics is still in its early stages and there are many other use cases to be explored. We expect that the development of blockchain-based solutions would change the current genomic ecosystem. As one of the aims of using blockchain is to empower patients to control their data, their role in data sharing will be significant. Increased trust and automated processes provided by blockchain and smart contracts would scale the amount of data being shared. In addition, blockchain allows the design of incentives that can facilitate sharing, storing, and processing genomic data in a fair way and for the ultimate purpose of advancing our knowledge of the human genome.

6 CONCLUSION

There is an increasing number of papers proposing blockchain based solutions to enable the storage, sharing and processing of genomic data. A similar trend can be observed from the large number of commercial/non-commercial blockchain applications that aim to enable genomic data exchange. In this paper, we provided a comprehensive overview of the existing efforts in this area.

Our study employed a taxonomy in which genomic applications of blockchain are classified into commercial, and non-commercial applications. Non-commercial applications have been further categorized according to their specific goals, namely data sharing, analysis, secure storage, access control, and logging/auditing. After providing certain details about each application, we described the advantages and drawbacks of the proposed approach and provided a comparison between the proposals concerning the blockchain platform selection, how data is stored, shared, and protected in the proposal. Software instability, interoperability, and the security risks associated with the rigidity of smart contracts are some of the highlighted challenges. Another challenge is protecting the privacy of the data and the identities of the data owner. Privacy-enhancing technologies such as those mentioned earlier can be beneficial when applied in these settings.

Our results suggest that immutability and decentralization are the main motivations for blockchain use in this context. We observed that empowering data owners to control their data is a common argument in the papers. In most papers, blockchain is used to give control of the data to organizations or individuals (patients).

We also observed that the applications or use-cases of blockchain in genomics are rather limited compared to financial applications although there is a huge potential in exploring alternative use-cases. We identified a list of open issues/challenges from the perspective of the existing proposals (Section 4.5). We recommend experimenting with blockchain-based distributed analytics (i.e. processing) in genomics. In addition, evaluating the performance of various privacy-enhancing technologies such as homomorphic encryption, multi-party computation, zero-knowledge proofs, and off-chain computation can shed some light on the feasibility of privacy-preserving distributed analytics in blockchain networks. Another promising future direction is exploring blockchain-based trusted and verifiable access to genomic data. The trust-over-ip principles [32] can be utilized to create and manage decentralized identities for researchers.

Our general impression is that genomic applications on blockchain are still in their early stages of research and development and require further transformation both socially and technologically in order to be adopted. Recent efforts to promote the use of blockchain in genomics such as [59] can help accelerate the adoption of this technology by showcasing the potential and feasibility of using it in various genomic applications.

REFERENCES

[1] 23andMe. 2021. *23andMe: DNA Genetic Testing & Analysis.* Retrieved 08.05.2021 from https://www.23andme.com

[2] Cornelius C Agbo, Qusay H Mahmoud, and J Mikael Eklund. 2019. Blockchain technology in healthcare: a systematic review. In *Healthcare*, Vol. 7. Multidisciplinary Digital Publishing Institute, 56.

[3] Eman Ahmed and Mahsa Shabani. 2019. DNA Data Marketplace: An analysis of the ethical concerns regarding the participation of the individuals. *Frontiers in genetics* 10 (2019), 1107.

[4] Mete Akgün, A Osman Bayrak, Bugra Ozer, and M Şamil Sağıroğlu. 2015. Privacy preserving processing of genomic data: A survey. *Journal of biomedical informatics* 56 (2015), 103–111.

[5] José Bacelar Almeida, Endre Bangerter, Manuel Barbosa, Stephan Krenn, Ahmad-Reza Sadeghi, and Thomas Schneider. 2010. A certifying compiler for zero-knowledge proofs of knowledge based on σ-protocols. In *European Symposium on Research in Computer Security*. Springer, 151–167.

[6] Joe Andrews. 2019. *23andMe competitor Veritas Genetics slashes price of whole genome sequencing 40% to $600.* Retrieved 12.02.2021 from https://www.cnbc.com/2019/07/01/for-600-veritas-genetics-sequences-6point4-billion-letters-of-your-dna.html

[7] Samuel J Aronson and Heidi L Rehm. 2015. Building the foundation for genomics in precision medicine. *Nature* 526, 7573 (2015), 336–342.

[8] Euan A Ashley. 2016. Towards precision medicine. *Nature Reviews Genetics* 17, 9 (2016), 507.

[9] Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013. Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society.* 95–106.

[10] Md Momin Al Aziz, Md Nazmus Sadat, Dima Alhadidi, Shuang Wang, Xiaoqian Jiang, Cheryl L Brown, and Noman Mohammed. 2019. Privacy-preserving techniques of genomic data—a survey. *Briefings in bioinformatics* 20, 3 (2019), 887–895.

[11] Leemon Baird. 2016. The swirlds hashgraph consensus algorithm: Fair, fast, byzantine fault tolerance. *Swirlds Tech Reports SWIRLDS-TR-2016-01, Tech. Rep* (2016).

[12] Rafael Belchior, André Vasconcelos, Sérgio Guerreiro, and Miguel Correia. 2020. A survey on blockchain interoperability: Past, present, and future trends. *arXiv preprint arXiv:2005.14282* (2020).

[13] Nazanin Zahed Benisi, Mehdi Aminian, and Bahman Javadi. 2020. Blockchain-based decentralized storage networks: A survey. *Journal of Network and Computer Applications* (2020), 102656.

[14] Bonnie Berger and Hyunghoon Cho. 2019. Emerging technologies towards enhancing privacy in genomic data sharing.

[15] Jorge Bernal Bernabe, Jose Luis Canovas, Jose L Hernandez-Ramos, Rafael Torres Moreno, and Antonio Skarmeta. 2019. Privacy-preserving solutions for Blockchain: review and challenges. *IEEE Access* 7 (2019), 164908–164940.

[16] Mikael Beyene, Scott Thiebes, and Ali Sunyaev. 2019. Multi-Stakeholder Consent Management in Genetic Testing: A Blockchain-Based Approach.

[17] Joseph Bonneau, Arvind Narayanan, Andrew Miller, Jeremy Clark, Joshua A Kroll, and Edward W Felten. 2014. Mixcoin: Anonymity for bitcoin with accountable mixes. In *International Conference on Financial Cryptography and Data Security*. Springer, 486–504.

[18] Christian Cachin et al. 2016. Architecture of the hyperledger blockchain fabric. In *Workshop on distributed cryptocurrencies and consensus ledgers*, Vol. 310. Chicago, IL.

[19] Davide Calvaresi, Daniel Cesarini, Paolo Sernani, Mauro Marinoni, Aldo Franco Dragoni, and Arnon Sturm. 2017. Exploring the ambient assisted living domain: a systematic review. *Journal of Ambient Intelligence and Humanized Computing* 8, 2 (2017), 239–257.

[20] Federico Carlini, Roberto Carlini, Stefano Dalla Palma, Remo Pareschi, Federico Zappone, and Daniele Albanese. 2020. The Genesy Model for a Blockchain-based Fair Ecosystem of Genomic Data. In *2020 Seventh International Conference on Software Defined Systems (SDS)*. IEEE, 183–189.

[21] Thomas Caskey. 2018. Precision medicine: functional advancements. *Annual review of medicine* 69 (2018), 1–18.

[22] David Chaum and Eugène Van Heyst. 1991. Group signatures. In *Workshop on the Theory and Application of of Cryptographic Techniques*. Springer, 257–265.

[23] Huashan Chen, Marcus Pendleton, Laurent Njilla, and Shouhuai Xu. 2020. A survey on ethereum systems security: Vulnerabilities, attacks, and defenses. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–43.

[24] clinicalgenome.org. 2020. *Clinical Genome*. Retrieved 2021-02-27 from https://clinicalgenome.org/

[25] European Commission. [n.d.]. *General Data Protection Regulation*. Retrieved 2016-04-27 from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

[26] U.S. Equal Employment Opportunity Commission. 2008. *Genetic Information and Nondiscrimination Act*. Retrieved 2016-04-27 from https://www.eeoc.gov/statutes/genetic-information-nondiscrimination-act-2008

[27] The Ethereum Community. 2018. *What is Ethereum?* Retrieved 05.11.2020 from https://github.com/ethereum/homestead-guide/blob/master/source/introduction/what-is-ethereum.rst

[28] Consensys. 2021. *Quorum blockchain platform*. https://consensys.net/quorum/

[29] Manuel Corpas, Nadezda V Kovalevskaya, Amanda McMurray, and Fiona GG Nielsen. 2018. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS computational biology* 14, 3 (2018), e1005873.

[30] Diana Crow. 2019. A new wave of genomics for all. *Cell* 177, 1 (2019), 5–7.

[31] S Mason Dambrot. 2018. ReGene: Blockchain backup of genome data and restoration of pre-engineered expressed phenotype. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 945–950.

[32] Matthew Davie, Dan Gisolfi, Daniel Hardman, John Jordan, Darrell O'Donnell, and Drummond Reed. 2019. The trust over ip stack. *IEEE Communications Standards Magazine* 3, 4 (2019), 46–51.

[33] Laura DeFrancesco and Ariel Klevecz. 2019. Your DNA broker. *Nature biotechnology* 37, 8 (2019), 842.

[34] Hanne Degans. 2019. *IMEC-Int News release - New genome analytics platform makes clinical genomics affordable for daily use in hospital*. Retrieved 2021-1-12 from https://www.imec-int.com/en/articles/new-genome-analytics-platform-makes-clinical-genomics-affordable-for-daily-use-in-hospital

[35] Preston Estep Mirza Cifric Yining Zhao Dennis Grishin, Kamal Obbad and George Church. 2018. *Nebula Genomics: Blockchain-enabled genomic data sharing and analysis platform*. Retrieved 01.12.2020 from http://arep.med.harvard.edu/pdf/Grishin_Church_v4.52_2018.pdf

[36] disgenet.org. 2020. *A database of gene-disease associations*. Retrieved 12.01.2021 from https://www.disgenet.org/

[37] dnatix.com. 2020. *DNAtix*. Retrieved 01.12.2020 from https://www.dnatix.com/

[38] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing. 2017. Manual for Using Homomorphic Encryption for Bioinformatics. *Proc. IEEE* 105, 3 (2017), 552–567. https://doi.org/10.1109/JPROC.2016.2622218

[39] Jacob Eberhardt and Stefan Tai. 2018. Zokrates-scalable privacy-preserving off-chain computations. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 1084–1091.

[40] embleema.com. 2020. *embleema*. Retrieved 01.12.2020 from https://embleema.com/

[41] encrypgen.com. 2018. *encrypgen*. https://encrypgen.com/wp-content/uploads/2018/12/Gene-ChainVer1.pdf

[42] Nikolaos Evangelatos, Sudhakara P Upadya, Julien Venne, Kapaettu Satyamoorthy, Helmut Brand, CS Ramashesha, and Angela Brand. 2020. Digital transformation and governance innovation for public biobanks and free/libre open source software using a blockchain technology. *OMICS: A Journal of Integrative Biology* 24, 5 (2020), 278–285.

[43] Exonum. 2020. *The Exonum platform*. Retrieved 01.02.2021 from https://exonum.com/index

[44] Marc Fiume, Miroslav Cupak, Stephen Keenan, Jordi Rambla, Sabela de la Torre, Stephanie OM Dyke, Anthony J Brookes, Knox Carey, David Lloyd, Peter Goodhand, et al. 2019. Federated discovery and sharing of genomic data using Beacons. *Nature biotechnology* 37, 3 (2019), 220–224.

[45] Greenspan G. 2015. *MultiChain Private Blockchain - White Paper*. Retrieved 05 from http://www.multichain.com/download/MultiChain-White-Paper.pdf

[46] ga4gh.org. 2020. *Global Alliance for Genomics and Health*. Retrieved 2021-02-27 from https://www.ga4gh.org/

[47] genecoin.me. 2020. *genecoin*. Retrieved 01.12.2020 from http://genecoin.me/

[48] Nature Genetics. 2019. In genetics, context matters. *Nature Genetics* 51, 10 (01 Oct 2019), 1425–1425. https://doi.org/10.1038/s41588-019-0515-7

[49] Rosario Gennaro, Craig Gentry, and Bryan Parno. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In *Annual Cryptology Conference*. Springer, 465–482.

[50] genomes.io. 2020. *genomes.io*. Retrieved 01.12.2020 from https://www.genomes.io/

[51] Mahdi Ghadamyari and Saeed Samet. 2019. Privacy-Preserving Statistical Analysis of Health Data Using Paillier Homomorphic Encryption and Permissioned Blockchain. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 5474–5479.

[52] Benjamin Scott Glicksberg, Shohei Burns, Rob Currie, Ann Griffin, Zhen Jane Wang, David Haussler, Theodore Goldstein, and Eric Collisson. 2020. Blockchain-Authenticated Sharing of Genomic and Clinical Outcomes Data of Patients With Cancer: A Prospective Cohort Study. *J Med Internet Res* 22, 3 (20 Mar 2020), e16810. https://doi.org/10.2196/16810

[53] Dennis Grishin, Kamal Obbad, Preston Estep, Kevin Quinn, Sarah Wait Zaranek, Alexander Wait Zaranek, Ward Vandewege, Tom Clegg, Nico César, Mirza Cifric, et al. 2018. Accelerating genomic data generation and facilitating genomic data access using decentralization, privacy-preserving technologies and equitable compensation. *Blockchain in Healthcare Today* (2018).

[54] Dennis Grishin, Jean Louis Raisaro, Juan Ramón Troncoso-Pastoriza, Kamal Obbad, Kevin Quinn, Mickaël Misbach, Jared Gollhardt, Joao Sa, Jacques Fellay, George M Church, et al. 2019. Citizen-Centered, Auditable, and Privacy-Preserving Population Genomics. *bioRxiv* (2019), 799999.

[55] Gamze Gürsoy, Robert Bjornson, Molly E Green, and Mark Gerstein. 2020. Using blockchain to log genome dataset access: efficient storage and query. *BMC medical genomics* 13, 7 (2020), 1–9.

[56] Gamze Gursoy, Charlotte Brannon, Sarah Wagner, and Mark Gerstein. 2020. Storing and analyzing a genome on a blockchain. *bioRxiv* (2020).

[57] Gamze Gürsoy, Charlotte M Brannon, and Mark Gerstein. 2020. Using Ethereum blockchain to store and query pharmacogenomics data via smart contracts. *BMC Medical Genomics* 13, 1 (2020), 1–11.

[58] Marko Hölbl, Marko Kompara, Aida Kamišalić, and Lili Nemec Zlatolas. 2018. A systematic review of the use of blockchain in healthcare. *Symmetry* 10, 10 (2018), 470.

[59] IDASH. 2021. *IDASH PRIVACY & SECURITY WORKSHOP.* http://www.humangenomeprivacy.org/

[60] Atalay M Ileri, Halil I Ozercan, Alper Gundogdu, Ahmet K Senol, M Yusuf Ozkaya, and Can Alkan. 2016. Coinami: a cryptocurrency with DNA sequence alignment as proof-of-work. *arXiv preprint arXiv:1602.03031* (2016).

[61] ipfs.io. 2020. *IPFS Powers the Distributed Web.* Retrieved 01.12.2020 from https://ipfs.io

[62] Vijayasri Iyer, AM Hima Vyshnavi, Sriram Iyer, and PK Krishnan Namboori. 2019. An AI driven Genomic Profiling System and Secure Data Sharing using DLT for cancer patients. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*. IEEE, 1–5.

[63] Xiao-Ling Jin, Miao Zhang, Zhongyun Zhou, and Xiaoyu Yu. 2019. Application of a Blockchain Platform to Manage and Secure Personal Genomic Data: A Case Study of LifeCODE. ai in China. *Journal of medical Internet research* 21, 9 (2019), e13587.

[64] Stephanie B Johnson, Ingrid Slade, Alberto Giubilini, and Mackenzie Graham. 2020. Rethinking the ethical principles of genomic medicine services. *European Journal of Human Genetics* 28, 2 (2020), 147–154.

[65] Harry Kalodner, Steven Goldfeder, Xiaoqi Chen, S Matthew Weinberg, and Edward W Felten. 2018. Arbitrum: Scalable, private smart contracts. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 1353–1370.

[66] Shafaq Naheed Khan, Faiza Loukil, Chirine Ghedira-Guegan, Elhadj Benkhelifa, and Anoud Bani-Hani. 2021. Blockchain smart contracts: Applications, challenges, and future trends. *Peer-to-peer Networking and Applications* (2021), 1–25.

[67] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology* 51, 1 (2009), 7–15.

[68] John Kolb, Moustafa AbdelBaky, Randy H Katz, and David E Culler. 2020. Core Concepts, Challenges, and Future Directions in Blockchain: A Centralized Tutorial. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–39.

[69] Tsung-Ting Kuo. 2020. The anatomy of a distributed predictive modeling framework: online learning, blockchain network, and consensus algorithm. *JAMIA open* 3, 2 (2020), 201–208.

[70] Tsung-Ting Kuo, Rodney A Gabriel, Krishna R Cidambi, and Lucila Ohno-Machado. 2020. EX pectation P ropagation LO gistic RE g R ession on permissioned block CHAIN (ExplorerChain): decentralized online healthcare/genomics predictive model learning. *Journal of the American Medical Informatics Association* 27, 5 (2020), 747–756.

[71] Tsung-Ting Kuo, Rodney A Gabriel, and Lucila Ohno-Machado. 2019. Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. *Journal of the American Medical Informatics Association* 26, 5 (2019), 392–403.

[72] Tsung-Ting Kuo, Xiaoqian Jiang, Haixu Tang, XiaoFeng Wang, Tyler Bath, Diyue Bu, Lei Wang, Arif Harmanci, Shaojie Zhang, Degui Zhi, et al. 2020. iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching.

[73] Tsung-Ting Kuo, Jihoon Kim, and Rodney A Gabriel. 2020. Privacy-preserving model learning on a blockchain network-of-networks. *Journal of the American Medical Informatics Association* 27, 3 (2020), 343–354.

[74] Tsung-Ting Kuo, Hugo Zavaleta Rojas, and Lucila Ohno-Machado. 2019. Comparison of blockchain platforms: a systematic review and healthcare examples. *Journal of the American Medical Informatics Association* 26, 5 (2019), 462–478.

[75] Laphou Lao, Zecheng Li, Songlin Hou, Bin Xiao, Songtao Guo, and Yuanyuan Yang. 2020. A survey of IoT applications in blockchain systems: Architecture, consensus, and traffic modeling. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–32.

[76] Peeter Laud and Alisa Pankova. 2018. Privacy-preserving record linkage in large databases using secure multiparty computation. *BMC medical genomics* 11, 4 (2018), 84.

[77] Katrina Learned, Ann Durbin, Robert Currie, Ellen Towle Kephart, Holly C Beale, Lauren M Sanders, Jacob Pfeil, Theodore C Goldstein, Sofie R Salama, David Haussler, et al. 2019. Barriers to accessing public cancer genomic data. *Scientific data* 6, 1 (2019), 1–7.

[78] Seo-Joon Lee, Gyoun-Yon Cho, Fumiaki Ikeno, and Tae-Ro Lee. 2018. BAQALC: blockchain applied lossless efficient transmission of DNA sequencing data for next generation medical informatics. *Applied Sciences* 8, 9 (2018), 1471.

[79] longenesis.co. 2020. *longenesis.* http://longenesis.com/

[80] lunadna.com. 2020. *Luna DNA.* Retrieved 01.12.2020 from https://www.lunadna.com/

[81] Jeantine E Lunshof, Ruth Chadwick, Daniel B Vorhaus, and George M Church. 2008. From genetic privacy to open consent. *Nature Reviews Genetics* 9, 5 (2008), 406–411.

[82] Shuaicheng Ma, Yang Cao, and Li Xiong. 2020. Efficient logging and querying for blockchain-based cross-site genomic dataset access audit. *BMC Medical Genomics* 13, 7 (2020), 1–13.

[83] Tim K Mackey, Tsung-Ting Kuo, Basker Gummadi, Kevin A Clauson, George Church, Dennis Grishin, Kamal Obbad, Robert Barkovich, and Maria Palombini. 2019. 'Fit-for-purpose?'–challenges and opportunities for applications of blockchain technology in the future of healthcare. *BMC medicine* 17, 1 (2019), 68.

[84] Nicholas Mamo, Gillian M Martin, Maria Desira, Bridget Ellul, and Jean-Paul Ebejer. 2019. Dwarna: a blockchain solution for dynamic consent in biobanking. *European Journal of Human Genetics* (2019), 1–18.

[85] Nicholas Mamo, Gillian M Martin, Maria Desira, Bridget Ellul, and Jean-Paul Ebejer. 2020. Dwarna: a blockchain solution for dynamic consent in biobanking. *European Journal of Human Genetics* 28, 5 (2020), 609–626.

[86] Garima Mathur, Anjana Pandey, and Sachin Goyal. 2020. Immutable DNA Sequence Data Transmission for Next Generation Bioinformatics Using Blockchain Technology. In *2nd International Conference on Data, Engineering and Applications (IDEA).* IEEE, 1–6.

[87] Felix Konstantin Maurer, Till Neudecker, and Martin Florian. 2017. Anonymous CoinJoin transactions with arbitrary values. In *2017 IEEE Trustcom/BigDataSE/ICESS.* IEEE, 522–529.

[88] Muhammad Izhar Mehar, Charles Louis Shier, Alana Giambattista, Elgar Gong, Gabrielle Fletcher, Ryan Sanayhie, Henry M Kim, and Marek Laskowski. 2019. Understanding a revolutionary and flawed grand experiment in blockchain: the DAO attack. *Journal of Cases on Information Technology (JCIT)* 21, 1 (2019), 19–32.

[89] Melinda C Mills and Charles Rahal. 2019. A scientometric review of genome-wide association studies. *Communications biology* 2, 1 (2019), 1–11.

[90] N Mofokeng and T Fatima. 2018. Future tourism trends: Utilizing non-fungible tokens to aid wildlife conservation. *African Journal of Hospitality, Tourism and Leisure* 7, 4 (2018).

[91] MPEG-G. 2021. *genecoin.* Retrieved 09.04.2021 from https://mpeg-g.org/

[92] Satoshi Nakamoto and A Bitcoin. 2008. A peer-to-peer electronic cash system. *Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf* (2008).

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

[94] Maurício Moreira Neto, Carlos Sérgio da S Marinho, Emanuel F Coutinho, Leonardo O Moreira, Javam de C Machado, and José Neuman de Souza. 2020. Research Opportunities for E-health Applications with DNA Sequence Data using Blockchain Technology. In *2020 IEEE International Conference on Software Architecture Companion (ICSA-C).* IEEE, 95–102.

[95] Sergey Popov Nikolay Kulemin and Alexey Gorbachev. 2017. *The Zenome Project: Whitepaper blockchain-based genomic ecosystem.* Retrieved 01.12.2020 from https://zenome.io/

[96] U.S. Department of Health & Human Services. 1996. *Health Insurance Portability and Accountability Act.* https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

[97] M. S. Ozdayi, M. Kantarcioglu, and B. Malin. 2020. Leveraging blockchain for immutable logging and querying across multiple sites. *BMC Med Genomics* 13, Suppl 7 (Jul 2020), 82.

[98] Halil Ibrahim Ozercan, Atalay Mert Ileri, Erman Ayday, and Can Alkan. 2018. Realizing the potential of blockchain technologies in genomics. *Genome research* 28, 9 (2018), 1255–1263.

[99] Raginee Pachaury and C Vasantha Lakshmi. 2021. Securing Genomics Data Using Blockchain Technology. In *Advances in Systems Engineering.* Springer, 473–480.

[100] Juha Partala, Tri Hong Nguyen, and Susanna Pirttikangas. 2020. Non-Interactive Zero-Knowledge for Blockchain: A Survey. *IEEE Access* 8 (2020), 227945–227961.

[101] Nicholas D Pattengale and Corey M Hudson. 2020. Decentralized genomics audit logging via permissioned blockchain ledgering. *BMC Medical Genomics* 13, 7 (2020), 1–9.

[102] PetaSuite. 2020. *PetaGene.* Retrieved 2021-02-27 from https://www.petagene.com/products/

[103] Mark Phillips. 2018. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Human genetics* 137, 8 (2018), 575–582.

[104] Drummond Reed, Manu Sporny, Dave Longley, Christopher Allen, Ryan Grant, Markus Sabadello, and Jonathan Holt. 2020. Decentralized identifiers (dids) v1. 0. *Draft Community Group Report* (2020).

[105] Ronald L Rivest, Adi Shamir, and Yael Tauman. 2001. How to leak a secret. In *International Conference on the Theory and Application of Cryptology and Information Security.* Springer, 552–565.

[106] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications* 10, 1 (2019), 1–9.

[107] Juan M Roman-Belmonte, Hortensia De la Corte-Rodriguez, and E Carlos Rodriguez-Merchan. 2018. How blockchain technology can change medicine. *Postgraduate medicine* 130, 4 (2018), 420–427.

[108] Tim Ruffing, Pedro Moreno-Sanchez, and Aniket Kate. 2014. Coinshuffle: Practical decentralized coin mixing for bitcoin. In *European Symposium on Research in Computer Security*. Springer, 345–364.

[109] Christoph Schickhardt, Henrike Fleischer, and Eva C Winkler. 2020. Do patients and research subjects have a right to receive their genomic raw data? An ethical and legal analysis. *BMC medical ethics* 21, 1 (2020), 1–12.

[110] Mahsa Shabani. 2019. Blockchain-based platforms for genomic data sharing: a de-centralized approach in response to the governance problems? *Journal of the American Medical Informatics Association* 26, 1 (2019), 76–80.

[111] Suyash S Shringarpure and Carlos D Bustamante. 2015. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics* 97, 5 (2015), 631–646.

[112] Khaled Shuaib, Heba Saleous, Nazar Zaki, and Fida Dankar. 2020. A Layered Blockchain Framework for Healthcare and Genomics. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 156–163.

[113] Vahan Simonyan and Raja Mazumder. 2014. High-performance integrated virtual environment (HIVE) tools and applications for big data analysis. *Genes* 5, 4 (2014), 957–981.

[114] Katerina Sotiraki, Esha Ghosh, and Hao Chen. 2020. Privately computing set-maximal matches in genomic data. *BMC Medical Genomics* 13, 7 (2020), 1–8.

[115] Manu Sporny, D Longley, and David Chadwick. 2019. Verifiable credentials data model 1.0. *W3C, W3C Candidate Recommendation, March* (2019).

[116] Stellar. 2020. *Stellar - An Open Network for Money*. Retrieved 01.12.2020 from https://www.stellar.org/

[117] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. 2015. Big data: astronomical or genomical? *PLoS biology* 13, 7 (2015), e1002195.

[118] Strand-NGS. 2021. *Strand NGS: Guide to Storage and Computation Requirements*. Retrieved 07.05.2021 from https://www.strand-ngs.com/support/ngs-data-storage-requirements

[119] Stripe. 2020. *Stripe - A Complete Payments Platform*. Retrieved 01.12.2020 from https://www.stripe.com

[120] Karim Sultan, Umar Ruhi, and Rubina Lakhani. 2018. Conceptualizing blockchains: characteristics & applications. *arXiv preprint arXiv:1806.03693* (2018).

[121] You Sun, Rui Zhang, Xin Wang, Kaiqiang Gao, and Ling Liu. 2018. A decentralizing attribute-based signature for healthcare blockchain. In *2018 27th International conference on computer communication and networks (ICCCN)*. IEEE, 1–9.

[122] Nick Szabo. 1997. Formalizing and securing relationships on public networks. *First monday* (1997).

[123] Asoke K Talukder, Manish Chaitanya, David Arnold, and Kouichi Sakurai. 2018. Proof of disease: A blockchain consensus protocol for accurate medical decisions and reducing the disease burden. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 257–262.

[124] Scott Thiebes, Niclas Kannengießer, Manuel Schmidt-Kraepelin, and Ali Sunyaev. 2020. Beyond Data Markets: Opportunities and Challenges for Distributed Ledger Technology in Genomics. In *Hawaii International Conference on System Sciences (Forthcoming)*.

[125] Scott Thiebes, Matthias Schlesner, Benedikt Brors, and Ali Sunyaev. 2019. Distributed Ledger Technology in genomics: a call for Europe. *European Journal of Human Genetics* (2019), 1–2.

[126] European Union. 2020. *Beyond 1M Genomes*. https://cordis.europa.eu/project/id/951724

[127] Daniel Uribe and Gisele Waters. 2020. Privacy Laws, Genomic Data and Non-Fungible Tokens. *The Journal of The British Blockchain Association* (2020), 13164.

[128] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[129] D. WSJ: Roland. 2019. *How Drug Companies Are Using Your DNA To Make New Medicine*. Retrieved 07.05.2021 from https://www.wsj.com/articles/23andme-glaxo-mine-dna-data-in-hunt-for-new-drugs-11563879881

[130] www.shivom.io. 2020. *shivom*. Retrieved 01.12.2020 from https://www.shivom.io/

[131] Yang Xiao, Ning Zhang, Wenjing Lou, and Y Thomas Hou. 2020. A survey of distributed consensus protocols for blockchain networks. *IEEE Communications Surveys & Tutorials* (2020).

[132] Xixun Yu, Zheng Yan, and Athanasios V Vasilakos. 2017. A survey of verifiable computation. *Mobile Networks and Applications* 22, 3 (2017), 438–453.

[133] Rui Zhang, Rui Xue, and Ling Liu. 2019. Security and privacy on blockchain. *ACM Computing Surveys (CSUR)* 52, 3 (2019), 1–34.

[134] Yihua Zhang, Marina Blanton, and Ghada Almashaqbeh. 2015. Secure distributed genome analysis for GWAS and sequence comparison computation. In *BMC medical informatics and decision making*, Vol. 15. 1–12.

[135] Yanjun Zhang, Xin Zhao, Xue Li, Mingyang Zhong, Caitlin Curtis, and Chen Chen. 2019. Enabling privacy-preserving sharing of genomic data for GWASs in decentralized networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 204–212.

[136] Shifa Zhangy, Anne Kim, Dianbo Liu, Sandeep C Nuckchadyy, Lauren Huangy, Aditya Masurkary, Jingwei Zhangy, Lawrence Pratheek Karnatiz, Laura Martinezx, Thomas Hardjono, et al. 2018. Genie: A Secure, Transparent Sharing and Services Platform for Genetic and Health Data. *arXiv*

preprint arXiv:1811.01431 (2018).

[137] Zibin Zheng, Shaoan Xie, Hong-Ning Dai, Xiangping Chen, and Huaimin Wang. 2018. Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services* 14, 4 (2018), 352–375.

[138] Hanrui Zhong, Yingpeng Sang, Yongchun Zhang, and Zhicheng Xi. 2019. Secure multi-party computation on blockchain: An overview. In *International Symposium on Parallel Architectures, Algorithms and Programming*. Springer, 452–460.

[139] Jiapeng Zhou, Yuxiang Feng, Zhenyu Wang, and Danyi Guo. 2021. Using Secure Multi-Party Computation to Protect Privacy on a Permissioned Blockchain. *Sensors* 21, 4 (2021), 1540.

[140] N Živi, E Kadušić, and K Kadušić. 2019. Directed Acyclic Graph as Tangle: an IoT Alternative to Blockchains. In *2019 27th Telecommunications Forum (TELFOR)*. IEEE, 1–3.

[141] Łukasz Żmudzin and Bartosz Sawicki. 2020. Design of Truly Distributed Storage for Large Medical Datasets. In *2020 IEEE 21st International Conference on Computational Problems of Electrical Engineering (CPEE)*. IEEE, 1–4.