Blockchain’s potential to address issues in genomics research and how it is being used today

Tony Jung and Richard Leu

ABSTRACT
Advancements in technology have greatly decreased the costs of genome sequencing and expedited the entire sequencing process. As a result, there has been a significant increase in the volume of genomic data. Although this is useful for genomics research, there are two major concerns with this increase in data. First, the greater volume of genomic data requires a substantial amount of computational resources to process and store this data. While cloud services can seem like an effective solution to process and store this data, cloud services aggregate their information in one data center which results in the risk of a single point of failure. With the increase in genomic data, there is also an increase in privacy concerns because genomic data contains personal and sensitive information. People are not comfortable with large companies that store genomic data and people do not want this data shared with the public. Blockchain is a network that can utilize numerous computers to process data and store multiple copies of the database to eliminate the risk of a single point of failure. The blockchain is also a decentralized network which means that it is not regulated by a third party. This allows the data contributors to have full ownership of their genomic data and give data contributors full control over this data.

AN OVERVIEW OF GENOMICS RESEARCH AND CURRENT ISSUES
Genomics is the field of research that studies the complete set of genetic material present in organisms. Within genomics, genome sequencing allows scientists to investigate the relationships between genetics and disease and can help identify the genetic mutations that underlie diseases and ultimately provide personalized diagnoses and treatment interventions. Advancements in DNA sequencing technologies have significantly decreased the cost of human genome sequencing, increasing the accessibility of genomic data to citizens at a price point under $1,000 USD. However, there are two issues associated with the increased volume of genomic sequencing data: First, the increased amount of genomic data requires substantial computational power and data storage. Second, genomic data contains sensitive information, and a growing load of genomic data consequently increases privacy concerns. This review will outline blockchain technology can solve these problems and provide an update on how blockchain is currently used in genomics research.

Present day genomics research utilizes a powerful technology called high throughput sequencing that can simultaneously sequence hundreds of millions of DNA molecules. Although useful, this technology produces an immense amount of data, requiring substantial computational power. As a result, computational analysis is the bottleneck of genomics research. A seemingly effective solution to address the increasing need for advanced computational power is to use cloud services. Cloud computing is a service that allows individuals to access and utilize additional computer processing power or data storage beyond their hardware’s capabilities from a shared supply of computational resources provided through the medium of the Internet. While this model is effective and scalable, aggregating data at one data center, as is the case with cloud services, increases the risk of a single point of failure. Security breaches of these data centers can result in the unauthorized access of genetic data and associated health records, as was the case in the breach of the DNA-testing firm Veritas Genetics in November 2019. This is a problem that cannot be simply fixed by constructing more data centers as they are ultimately difficult to both develop and maintain. The utilization of blockchain alternatives can not only increase the security of these sensitive data networks, but also incentivize safe commercialization practices of genetic data between data providers and buyers, as will be illustrated in the next sections of this review following an introduction of the intricacies underlying cryptocurrencies and blockchain.

INTRODUCTION TO CRYPTOCURRENCY AND BLOCKCHAIN
Cryptocurrencies such as bitcoin are digital currencies that have become increasingly popular since 2009. There are three major properties of cryptocurrencies that differentiate them from other digital and fiat currencies such as the United States Dollar. First, cryptocurrencies are decentralized meaning they are not controlled by a central authority; this marks a clear distinction from the aforementioned fiat currencies regulated by governments. Furthermore, all cryptocurrency transaction records, and other relevant data are recorded and immutable making it difficult to fabricate records. Unlike traditional centralized databases, transactions and records on the blockchain require a consensus from all participants on the network, thus rendering attempts at data fabrication much more costly and difficult to coordinate. Lastly, cryptocurrency users benefit from enhanced security in that their accounts are protected by cryptographic encryption.

Cryptocurrencies are traded on a public network called the blockchain. The data record on a blockchain, also known as the ledger, is shared across multiple public or private computers so that multiple copies of the database exist. These multiple computers are known as “nodes”, which support the blockchain network by
comparing and verifying ledger records with one another to verify the legitimacy of transaction records.9 Ledgers are ultimately composed of multiple data “blocks” that contain transaction records and other data. These blocks are created through a process called “mining” where miners—individuals who embed data onto the blockchain network—utilize computational power to package data into blocks, which are then verified by all nodes of the network.7 In exchange for the successfully mined block, miners are rewarded with cryptocurrency.7 This process creates a data “chain” in which each block corresponds to both the preceding and proceeding blocks.9 Immutability of the blockchain is such that only future blocks at the time of creation can be altered, whether it be in the mining or verification process—in other words, the nodes supporting the blockchain network would recognize any changes in the ledger such as attempts to manipulate a previous block as an entirely different network.11 Ultimately, these properties provide the framework necessary for a reliable, decentralized network that can securely store data and transaction records.7 While cryptocurrencies serve a useful medium of exchange, the blockchain technology itself has potential applications in other fields such as education, healthcare, and research.9

INCORPORATION OF BLOCKCHAIN TECHNOLOGY INTO GENOMICS

Blockchain technology can address both the computational processing demands and the storage requirements for large volumes of data that genomic databases increasingly require.4 Currently, the computational power used to mine and secure the bitcoin blockchain is approximately 126 quintillion hashes per second,12 - greater than the computing power of the Berkeley Open Infrastructure for Network Computing, the most powerful scientific computation network in the world.3 The large network of computers constituting a blockchain may be similarly used to process, exchange, and store genomic data. Miners can be incentivized by cryptocurrency rewards to contribute their computational resources for this purpose.4 Within the blockchain, copies of encrypted genomic data would be distributed across the multiple nodes which eliminates the risk of a single point of failure.1

Given that genomic data contains personal and sensitive information, people naturally have privacy concerns with companies that collect it.3 The establishment of a public blockchain to store this genomic data can address these issues. A public blockchain is decentralized which means that the data in this blockchain is not owned or regulated by a third party.5 Instead, a public blockchain gives contributors of genomic data ownership of their own data.4 However, while the data on public blockchains can be accessed during the verification process by nodes of the network, these blockchains can utilize homomorphic encryption to encrypt the data and thus maintain the privacy of data contributors.5 Homomorphic encryption is a cryptographica encryption method that allows for the processing of already encrypted data directly to generate further encrypted data.11 This differs from conventional encryptions which require decryption of the data before it can be analyzed.12 Because the encrypted data would not be decrypted before or after processing, homomorphic encryption would allow miners to use their computational resources to process genomic data while preserving the privacy of the data contributors. The data owners can then obtain the decrypted data and choose to distribute it to interested buyers.9

CURRENT BLOCKCHAIN TECHNOLOGICAL APPLICATIONS IN GENOMICS

There are currently several companies that use blockchain technology to store genomic data. One such company, genomes.io, is a DNA and health data bank that allows users to have their genomes sequenced and stored on their own platform.14 Through homomorphic encryption, only the data provider has access to their own data.4 Data providers can choose to anonymously grant certain buyers permission to view their data in exchange for GENE tokens that can be exchanged for fiat currency.14 Luna DNA, another company, offers a similar platform but is unique in that they distribute their company’s shares to data contributors.13 allowing users to share in the profits gained from research on Luna DNA’s database.14 Nebula Genomics is another company that utilizes the blockchain to give people control of their own genomic data.10 The founders of Nebula Genomics state that mainly privacy and cost deter people from sequencing their genome.9 With the blockchain and homomorphic encryption, Nebula Genomics, like genomes.io, allows data owners to anonymously and securely share their data in exchange for financial rewards.10 To address the cost concern of the genome sequencing, Nebula sequences one’s genome for as low as $99.16 Moreover, Nebula Genomics offers some customers free sequencing services thanks to partnerships with companies such as EMD Serono.16 EMD Serono is using the data on Nebula Genomic’s blockchain to further advance their lung cancer research.16 This partnership can provide EMD Serono valuable information and can help researchers accelerate medical developments.16 DNatix also offers a blockchain that can anonymously and securely store and sell genomics data.17 It is the first company to upload DNA sequences on a blockchain and uses a proprietary algorithm that can significantly compress data (99% compression rate).17 Such algorithms that can compress genomic data can effectively address the data storage concerns associated with a growing amount of genomic data.

Authenticated Resources in a Hashed Block Registry (ARBoR) is an application that utilizes the features of a blockchain to address another problem in genomics. Clinical genome sequencing is useful for the diagnosis and treatment of various illnesses.11 However, continuous advancements in genetics research require constant reanalysis and consequent updating of existing clinical genomic data reports.11 With such frequent revisions, there are several concerns that arise: For example, it may be difficult to keep track of the changes and ensure that only the most updated clinical genomic reports are distributed to clinicians and patients.11 Furthermore, records may be maliciously altered without detection.11 ARBoR solves these potential issues by storing the data of each single report as a single data block in the ARBoR ledger.15 When a report is revised, a new block is specifically made to fit and link to the block that contains the data of the previous version of the report.11 This process ultimately creates a chronological history of record revisions that is easy to validate but difficult to maliciously alter.8
CHALLENGES OF IMPLEMENTING BLOCKCHAIN

Although there are potential benefits of blockchain for genomic research, widespread adoption of this technology faces several barriers. One of the biggest challenges may stem from the association between cryptocurrencies and crime. For example, malicious hackers demand ransom payments in cryptocurrency form, and illicit drugs are purchased online with cryptocurrency. Such events tarnish the reputation of blockchain technology. Another challenge is that blockchain and cryptography are complex and difficult to understand, and people may not be open to working with the blockchain for this reason. Lastly, the blockchain is not environmentally sustainable as it requires enormous amounts of electricity. However, the blockchain is still an early technology that can be further developed and improved, with research being done to produce blockchains that require less computational power to decrease electricity usage.

CONCLUSION

Blockchain is a new and exciting technology that can address the needs for substantial computational resources and reliable privacy features for genomic data. With the blockchain, numerous nodes may be used to store and process large volumes of genomic data without having a single point of failure. Several companies have realized this potential and adopted the blockchain to serve this purpose. Apart from genomics, there are many other potential applications of the blockchain and the world should take full advantage of this technology.

REFERENCES


