

Realizing the potential of blockchain technologies in genomics

Halil Ibrahim Ozercan,¹ Atalay Mert Ileri,² Erman Ayday,¹ and Can Alkan¹

¹Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey; ²Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Genomics data introduce a substantial computational burden as well as data privacy and ownership issues. Data sets generated by high-throughput sequencing platforms require immense amounts of computational resources to align to reference genomes and to call and annotate genomic variants. This problem is even more pronounced if reanalysis is needed for new versions of reference genomes, which may impose high loads to existing computational infrastructures. Additionally, after the compute-intensive analyses are completed, the results are either kept in centralized repositories with access control, or distributed among stakeholders using standard file transfer protocols. This imposes two main problems: (1) Centralized servers become gatekeepers of the data, essentially acting as an unnecessary mediator between the actual data owners and data users; and (2) servers may create single points of failure both in terms of service availability and data privacy. Therefore, there is a need for secure and decentralized platforms for data distribution with user-level data governance. A new technology, blockchain, may help ameliorate some of these problems. In broad terms, the blockchain technology enables decentralized, immutable, incorruptible public ledgers. In this Perspective, we aim to introduce current developments toward using blockchain to address several problems in omics, and to provide an outlook of possible future implications of the blockchain technology to life sciences.

Following the scientific breakthroughs over several centuries, we are now in the era of “big data,” in which a significant portion of science and knowledge discovery relies on efficient processing of very large-scale data sets. There are many big data application domains, from astrophysics to targeted marketing and advertising, quantum physics, and the topic of this Perspective: life sciences, especially genomics (Stephens et al. 2015). Within the realm of big data, computational problems manifest themselves in data acquisition, storage, distribution, and analysis.

Bioinformatics challenges associated with genomics, or any other omics field, include compute-intensive data analysis and privacy-aware data storage and sharing. In today’s genomics research, analysis of high-throughput sequencing (HTS) data is common practice, which involves several computationally demanding steps such as read mapping and variation calling. Other applications of biological data analysis, such as calculating tertiary structures of RNA and protein molecules, are also computationally infeasible; therefore, they require immense amounts of computation (i.e., NP-complete) (Hart and Istrail 1997; Akutsu 2000).

Contemporary genomics research deals with two problems regarding data storage and sharing. First, the privacy of the individuals who contribute biological material such as DNA should be preserved. The second issue deals with the question: “Who controls the data?” Ideally, the sample provider (i.e., the patient) should be able to control access either directly or through trusted third parties, such as their doctors or research groups with necessary permissions and ethical board approvals. Currently such control is possible through centralized data repositories; however, both granting and revoking access to data usually takes long processing times.

In this Perspective, we focus on computation and privacy-aware data sharing problems in genomics, and potential use of

the blockchain technology in addressing some of the computational problems. We note that this paper aims to review only the applications of blockchain in genomics, and not the technology itself. We refer the interested reader to several other reviews on blockchain technology (Tapscott and Tapscott 2016; Witte 2016; Miraz and Ali 2018).

Genomic “big data”

High-throughput DNA sequencing (HTS) technologies evolved very quickly in the last decade, and now they are among the most powerful tools available for biological research (Metzker 2010). We are now able to read the entire genome of a human individual in a few days for a fraction of the costs incurred by previous technologies (Metzker 2010; Goodwin et al. 2016). However, the volume of data generated by these platforms is enormous, leading to a picture in which computational analyses represent the major bottleneck (Flicek 2009; Sboner et al. 2011; Treangen and Salzberg 2012). For example, the Illumina HiSeq X Ten platform is reported to be able to sequence the genomes of approximately 18,000 humans a year, at an estimated cost of ~\$1200 per genome (<http://www.illumina.com/systems/hiseq-x-sequencing-system>.ilmn). This corresponds to ~2 petabytes of data per year, per sequencing center. Considering that there are many genome centers that either already have purchased or will purchase this system, the amount of data generated each year will increase from hundreds of petabytes to exabytes.

In light of this big data revolution in genomics, modern solutions lean toward utilizing professional infrastructures that can take such loads. Cloud services are very powerful in terms of both scalability and usability from a researcher’s perspective. However, cloud architectures gather all resources into one data center and therefore create a potential single point of failure.

Corresponding author: calkan@cs.bilkent.edu.tr

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.207464.116>. Freely available online through the *Genome Research* Open Access option.

© 2018 Ozercan et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Possible failures include not only infrastructure outage, but also data leaks and accesses to raw data by malicious parties. Although trade-off between privacy and fast analysis usually favored the latter until recently, increasing privacy concerns among individuals started to shift the tide (Wang et al. 2017). On the other hand, the alternative decentralized methods lacked a scheme to reach a consensus between peers on data ownership and access control. While other problems such as network bandwidth and privacy remain unsolved, blockchain offers a simple solution to consensus issues. Lately, blockchain-based approaches in genomics often utilize blockchain as a decentralized database medium. However, like many new technologies, there is a growing hype around blockchain. Therefore, it is important to realize that blockchain is only a tool with limitations that may help solve some problems. Nonetheless, overexpectations and concerns should not devalue the potential of this new technology that might be of importance in decentralized scientific computation.

A primer for the blockchain technology and cryptocurrencies

Blockchain, in a broad sense, is a distributed and immutable database, shared and automatically synchronized among all participants (Tapscott and Tapscott 2016). This distributed database technology was first developed to be used as a public ledger in the popular decentralized cryptocurrency, Bitcoin (Nakamoto 2008). Although Bitcoin and other similar cryptocurrencies were introduced as decentralized alternatives to coinage and monetary

systems that virtually remained unchanged since the time of the ancient Lydians, they are in fact mere applications of the underlying blockchain technology. The most important aspects of blockchain technology are: (1) decentralization (i.e., a single entity cannot control the database), (2) immutability (i.e., no past record can be altered), and (3) security (i.e., accounts are protected by enhanced cryptographic methods).

Decentralization is the essential contribution of blockchain to modern consensus agreements like legislation, financial agreements, or joint resolution. Most current approaches require a third party, a governor, to reach and force an agreement between members. This central authority should be trusted by all participants to fulfill arrangement conditions. However, additional measures are required to keep the authority in check in regard to potential abuse of power. These measures are likely to only increase in numbers and introduce many more actors that would clutter the system. Blockchain mitigates this trust to an algorithmic process. Every member can inspect the actions of others in a timely and organized fashion, which facilitates reaching a consensus at any given time. We provide the details of how blockchain achieves this in Box 1 for the interested reader.

Although it was developed as an integral part to Bitcoin, the blockchain technology itself is loosely coupled to Bitcoin and other cryptocurrencies as described above, making it possible to be used in other cases. Almost all applications of the blockchain technology can be boiled down to two interconnected processes called “Mining” and “Transaction” (Akcora et al. 2017). In a nutshell, mining refers to the creation of new blocks, that are “chained” one after another to form the blockchain, whereas transaction

Box 1. Inner workings of blockchain

Basic cryptography. Before explaining how blockchain handles immutability and security, we need to introduce the basics of modern cryptographic methods. The most known cryptographic method that uses the same key for both encryption and decryption is called symmetric key encryption. However, this method needs a secure communication channel to transfer the key in the first place. Public Key Cryptography (PKC) (Diffie and Hellman 1976) deals with this problem by offering asymmetric key pairs. The sender encrypts the message by using the intended recipient’s public key, which is accessible by everyone. The recipient then uses their corresponding private key to decrypt the message. This cryptographic model is highly utilized in security protocols due to its solid mathematical background. The most important use cases of PKC are digital signatures. One can prove the authenticity of a piece of data by generating a signature using their private key and then the signature could be verified by public key.

Cryptographic hashes are summaries of data in binary format in which one small change in the original data yields a 50% chance of changing every bit of the earlier hash value. This means that it is impossible to find data that corresponds to a desired hash value due to its highly probabilistic and volatile nature. On the contrary, it is effortless to generate the hash of a given piece of data.

Homomorphic encryption. Most blockchain approaches for sensitive data aim for access control and protecting the integrity of data (e.g., by keeping logs). For sensitive data types, privacy of the shared data becomes important. One way to protect the privacy of data is encryption. However, traditional encryption techniques require users to decrypt the data in order to operate on it (which is not desirable due to privacy concerns).

Homomorphic encryption enables computing on encrypted data without having to decrypt it. Fully homomorphic encryption (FHE) (Gentry 2009) allows conducting all operations on encrypted data; however, it is not practical for real-life implementation. Due to this practicality issue, variations of FHE, such as partially homomorphic encryption and somewhat homomorphic encryption, have emerged. Such encryption techniques only allow limited types (or a limited number) of operations on encrypted data, but they are shown to be practical for real-life implementation. For instance, Ayday et al. (2013) used partial homomorphic encryption to conduct personalized genomic testing on encrypted VCF files. Similarly, Yasuda et al. (2013) used somewhat homomorphic encryption for privacy-preserving DNA pattern matching.

Building consensus through immutability. Although the members agree on the current resolution of events, someone might claim that they actually did not commit a previous action which is now part of the consensus, e.g., claiming that a purchase was not conducted after receiving the item and demanding to get the funds back. Blockchain offers immutability to prevent such claims. Nonetheless, immutability is a double-edged sword in the sense that theft cannot be recovered. If malicious users wish to remove an earlier record from the chain, they have to go back in time to the block that hosts the record and start mining new blocks from there and catch up with the network. As long as malicious users do not hold >51% of computational power (ability to generate new blocks), they cannot catch up with the network and will fail to make an attack. Generating new blocks, also known as proof-of-work, is based upon finding a hash between a range (Box 2). As we described above, finding the required hash is completely random and demands random trials.

Security of the individual accounts is guaranteed by PKC. User identities are defined by wallets which are pairs of public and private keys. From the public key, an address that also provides some level of anonymity is generated to receive payments. Whenever someone wants to transfer funds, they prepare a receipt and sign it with their private key, which can be verified by anyone who has the public key. This handles the authenticity of transactions.

refers to records that are included in blocks, stating an exchange of assets between users or a state transition in the same chain. We provide more details and definitions of blockchain-related terms in Box 2, and we also introduce a simple analogy to further explain how blockchain functions in Box 3.

In the case of abstract blockchain technology, transactions refer to a state transition rather than actual exchange of assets (Buterin 2014). A transaction can include a piece of data to store or execute a term in a contract. This way, blockchain can be more than a central bank ledger but an abstract database of time-stamped data.

Mining is purposefully slowed down to an average constant time, which is called “dynamic difficulty level” to regulate block generation. The idea behind this slowdown is that if a malicious party intends to break the consensus or alter the blockchain, they are required to prove that they invested a greater amount of work than did the rest of the network in a much shorter time. Also, the economic integrity of the currency is kept in check by adjusting coin generation in cryptocurrencies, similar to limiting banknote production in national mints. Slowdown is achieved by either a computationally difficult process called “proof-of-work” (PoW), or a generic algorithm called “proof-of-stake” (PoS) that relays the blocks by a schedule based on accounts’ balance (Box 2; Tapscott and Tapscott 2016). However, we emphasize that, although Bitcoin was the first successful blockchain implementation, it and other cryptocurrencies remain as merely financial applications of the blockchain technology.

After blockchain technology gained attention due to Bitcoin, its broader potential was unleashed by the Ethereum Project proposed in late 2013 (Buterin 2014). Ethereum revolutionized the blockchain by adding a new type of autonomous account, called “Smart Contracts,” which enabled custom application develop-

ment on blockchain. This is achieved by providing a platform where a program could compute any set of instructions defined in a “smart contract.” Briefly, smart contracts include terms of agreements (i.e., work to be performed, resources to be allocated, etc.), and reward and penalty mechanisms when the agreements are met or unmet, respectively (Fig. 1). Through this generalization mechanism provided by Ethereum, the need to develop a new blockchain for each different purpose is eliminated. Current possible use cases of Ethereum include decentralized data feeds, cloud computing, prediction markets, and decentralized file storage (Raval 2016).

As we describe above, blockchain technology is still in the development stage where problems such as scalability (Croman et al. 2016; BigChainDB GmbH 2017; J Teutsch and C Reitwiessner, unpubl.), performance (Scherer 2017; Spasovski and Eklund 2017), and security (McCorry et al. 2016; Sapirshtein et al. 2016) are being addressed. Also, blockchain utilizes a wide range of cryptographic and computational tools to solve trust problems and build consensus. Expertise in cryptography may be required to fully understand the underlying technology; however, a number of reviews are available to provide basic information for the general reader (Bonneau et al. 2015; Crosby et al. 2016).

Use of blockchains in life sciences

Properties of the blockchain technology make it useful to address several problems in life sciences. Below, we explain how to leverage blockchains to reward resource-sharing (both computation and storage), facilitate decentralized data distribution, promote collaborative work, and provide genome privacy. We summarize the current state of blockchain usage in life sciences and other possible use cases in Table 1.

Box 2. Definitions of terms related to blockchain

Mining. To ensure that blockchain is very difficult to be manipulated, a sufficient amount of work is expected to be performed to create a new block. The basic requirement of this task is to be difficult to compute but easy to verify. Although it is time- and resource-consuming, block generation is vital for blockchain to be functional. Therefore, any person who puts resources into this task gets rewarded as incentive. This process is similar to mining precious metals, in which finding the material is based partially on luck, when found it is trivial to understand if the material is non-fake, and more manpower makes it easier to dig large areas.

Proof-of-work. A proof-of-work is a piece of data that is difficult (costly, time-consuming) to produce but easy for others to verify and which satisfies certain requirements. Producing a proof-of-work can be a random process with low probability so that much trial and error is required on average before a valid proof-of-work is generated.

Proof-of-stake. A proof-of-stake is a consensus algorithm like proof-of-work that decides on who mines the next block. The major difference is that proof-of-stake does not require vast computational power. Thus, it eliminates the need for large electricity consumption. The higher the stake someone puts in as deposit, the higher they earn from transaction fees.

Proof-of-space. Instead of proving a capability in terms of computation, proof-of-space utilizes memory-bound functions. This eliminates large electricity consumption of CPU-bounded functions while increasing the demand for larger space. There is also another approach in which users send files to each other and show a proof that the file is stored on the other end.

Transaction. Every transaction can be considered as state-transition-function. Blockchain starts in an empty state. A transaction moves a piece of data (coins) from an address to a new address if (1) the sender account has a sufficient amount, and (2) the transaction issuer proves to be the real owner of the sender address. If this state-transition-function returns “True,” then the transaction is considered to be valid and added to a candidate block by miners.

Difficulty level. The difficulty stems from the proof-of-work, which entails finding a number called nonce by pure chance. It cannot be calculated but can be found by trial and error. Therefore, a higher trial capability increases the chance of finding a valid nonce. The difficulty level gets updated at every 2016 blocks according to how long it took to generate the last 2016 blocks, which is around 2 wk in the Bitcoin network. This effectively limits the amount of new blocks generated by the miners, preventing devaluation of the money as more blocks are mined.

Smart contracts. Briefly, smart contracts are sets of instructions that are enforced when certain conditions are met, and whose authenticity, conditions, and necessities can be observed and approved by everyone. A smart contract operates as an autonomous account on the blockchain. It has a dedicated storage to keep details, objects, and information related to its application. Transactions that are addressed to a smart contract cause an activation and the contract updates the records depending upon its predefined instructions.

Box 3. A simple analogy for blockchain use

Imagine a town where, election after election, every mayor fails to solve the problem of immense corruption. All failing banks start to take advantage of the situation and blame the inadequate government for the economic crisis. Finally, the townspeople decide that a central authority cannot keep them safe anymore. A citizen proposes a solution that is described as follows:

A town meeting is called, and everyone joins with a new notebook.

All citizens report and prove how much money they own. Everyone takes note of each other.

After the meeting, when someone makes a transaction for any reason, they must announce this to everyone they know.

If someone hears about a transaction, they make a note of this transaction to the ongoing page of their notebook. They also must pass the information until it eventually reaches all townspeople.

The transaction is considered to be completed, and accounts are updated when the page that contains it is closed.

Everyone must close an ongoing page roughly at the same time. To achieve this without any centralized involvement, scientists of the town proposed a self-updating puzzle. When someone solves this puzzle, they publish their result with their ongoing page. If the solution is correct and their ongoing page does not include a faulty transaction, everyone copies the given page and closes it after adding a "rewarding transaction" (i.e., new "money") to the solver.

In this example, the virtual town uses the blockchain technology. Solving a puzzle refers to mining, while notebooks are the copies of the blockchain. Each page is a block, and the current state of the accounts can be inspected by checking this notebook from the start to the last closed page. People announcing and delivering transactions to one another constitutes the peer-to-peer network structure of the blockchain.

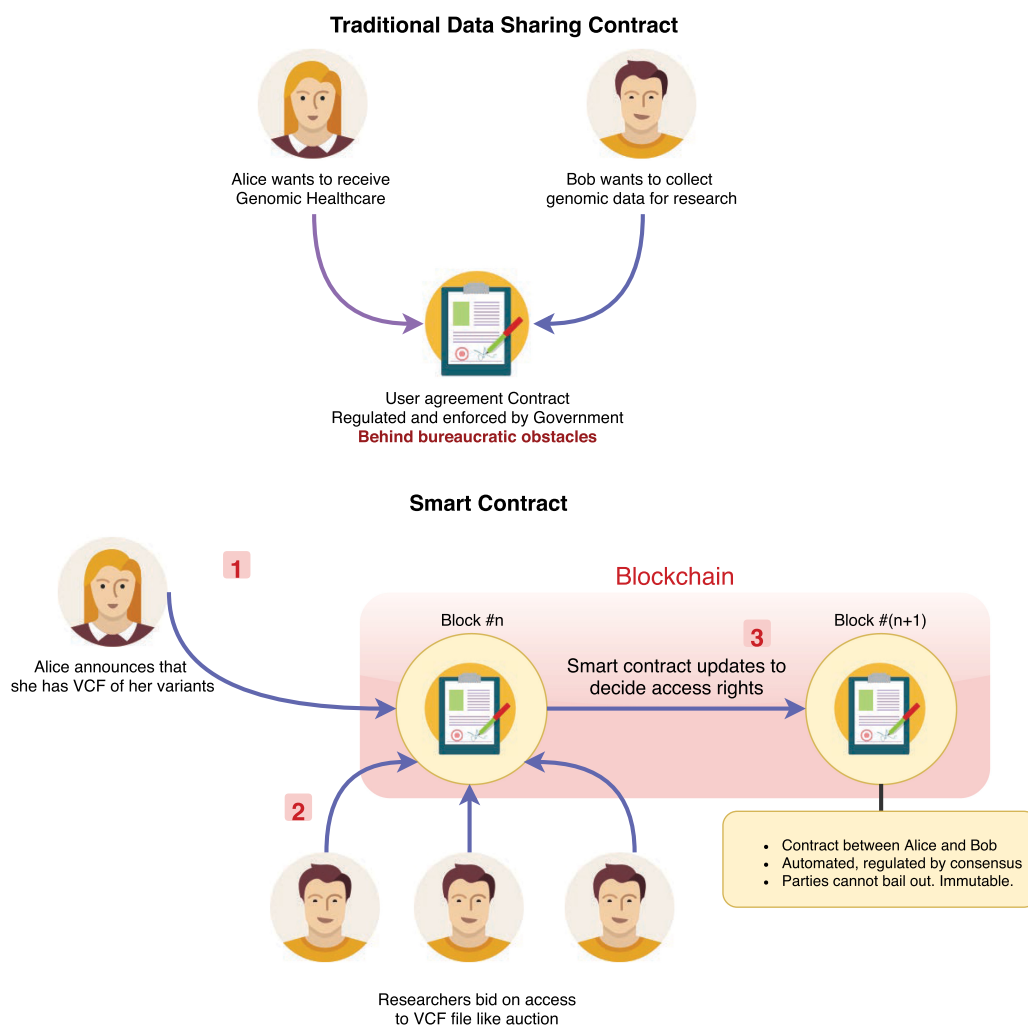


Figure 1. A smart contract to manage access management for genomic data. In traditional agreements, both parties sign a contract that dictates the boundaries, which the participants must obey. A third party is often required to enforce the agreement conditions. Smart contracts, on the other hand, can eliminate the need for a third party. In this figure, Alice publishes an encrypted version of her VCF file. At first, no other participant or researcher can analyze this file. In the second round, smart contract accepts bidding transactions for this file. The highest bidder is then selected to be the rightful owner and gets access to the file through an algorithmic process.

Table 1. Possible use cases for blockchain in genomics

| Use case | Examples |
|---|--|
| Distributed computation | Coinami, Gridcoin, Curecoin, FoldingCoin, Single instruction multiple data parallelization (smart contracts working on different data instances) |
| Data storage and distribution | Filecoin, CGT, CrypDist, Gene-chain, Nebula Genomics |
| Voting | Standardization teams (vote on proposals), Crowdsourced solutions (manual curation, gene-drug interactions) |
| Identity and ownership | Decentralized researcher identification databases, personal data adoption, crediting data ownership |
| Decentralized Autonomous Organization (DAO) | Predefined rules governing large organizations and projects such as GA4GH, ELIXIR, TCGA, ICGC |

Here we show several use cases of the blockchain technology with matching proposal projects or potential usage fields. (1) For distributed computation, there are already ongoing projects that utilize a blockchain for rewarding such as Gridcoin, Curecoin, and FoldingCoin. Coinami also operates job distribution on top of a blockchain. Most of these projects are built on the idea that Single Instruction Multiple Data (SIMD) frameworks can be distributed to volunteers on blockchain as long as data privacy is ensured. (2) Distributed data storage that is facilitated by blockchain is also raising attention due to high costs associated with cloud platforms. Filecoin (<https://filecoin.io/>) is an operational example of decentralized network storage. There are also many proposals to provide a free marketplace for private data while giving individuals the full control such as Nebula Genomics and Gene-chain. (3) Blockchain also provides a medium for secure online voting. Genomics research heavily relies on standardization, which is decided by a voting process. Additionally, crowdsourcing attempts, such as variant curation, can be implemented through blockchain that can also incorporate multiuser agreement on the curation results. (4) Identity of individuals (e.g., ORCID database) and ownership of data can be validated through blockchain. (5) Finally, Decentralized Autonomous Organizations such as Global Alliance for Genomics and Health (GA4GH), European life science infrastructure for biological information (ELIXIR), The Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium (ICGC) may use smart contracts as operating mediums to predefine rules, regulations, and governance.

Cryptocurrency system to reward compute-intensive analyses

Large computational clusters and the cloud platforms are often the “weapon of choice” in dealing with compute-intensive data processing problems such as read mapping. However, installation and maintenance of large data centers (including physical setups for cloud platforms) are difficult tasks by themselves. Additionally, in life sciences, the algorithms and reference databases are under constant development, and existing data typically need to be reanalyzed with the newest tools and reference databases in addition to the newly generated data. This never-ending cat-and-mouse game between the computational supply and demand necessitates alternative approaches to increase computing capabilities.

Another approach for dealing with compute-intensive applications is utilizing computational grids similar to the network infrastructures of cryptocurrencies. As of May 2018, Bitcoin mining network’s total computation power has reached over 33,900 PetaHash/s (<http://blockchain.info/stats>).³ In comparison, the world’s most powerful scientific computation grid (Berkeley Open Infrastructure for Network Computing [BOINC]) boasts a computation power of 19 PetaFLOP/s (<http://boinc.berkeley.edu/>).⁴ Although two units are not directly comparable because hashing uses integer operations, whereas BOINC’s power is measured in floating point operations, even the arbitrary and extremely pessimistic scaling of 1 to 1750 (assuming floating point operations are 1750× “harder” than integer operations) posits Bitcoin network more powerful than the BOINC network.

The computation power of the Bitcoin and similar cryptocurrency networks is solely used to maintain the currency’s integrity by ensuring that the new block creation is always a difficult task through proof-of-work (Box 2). The proof-of-work schemes within

Bitcoin and other cryptocurrencies are effective solutions for the security and financial integrity of the cryptocurrency systems, yet serve no other practical purpose. In fact, according to estimates, the Bitcoin network consumes more electricity than Denmark as of December 2017 (Digiconomist 2017). Although replacing wasteful computation with ecofriendly alternatives is an ongoing research area, most of them focus on the integrity aspect of proof-of-work in which the alternative solution must satisfy all the requirements that are provided by proof-of-work. On the other hand, mining is an incentivization scheme in which the miners only care about receiving the final reward in cryptocurrency form. Therefore, it may be possible to offer another reward mechanism that utilizes compute-heavy tasks in omics. Below, we outline some of the current attempts that try to accomplish this using different approaches, including token distribution and proof-of-stake (Box 2).

The Coinami project was recently proposed as a prototype volunteer grid computation platform with cryptocurrency awards to distribute the HTS read mapping work load to many volunteers (or, miners) and then collect and validate the results (Ileri et al. 2016). Most current cryptocurrencies, including Bitcoin, are completely decentralized, as the assignments for the proof-of-work can be generated independently by the miners as long as they meet the system’s difficulty level, which is determined by history of the blockchain. However, for the proof-of-work in Coinami to be useful, the assignments need to have real and practical value. Therefore, unlike major cryptocurrencies, Coinami is not completely decentralized due to the need for availability and generation of HTS data. Instead, Coinami has a federated structure, in which one root authority tracks and validates middle-level subauthority servers that supply HTS data to the system and checks for validity of alignments, and the third level is composed of miners (Fig. 2).

The root authority is trusted by the entire system to validate only “trustable” authority servers, which might be major sequencing centers. The root authority assigns certificates to the middle level authorities, validating and granting them permission to operate within the Coinami network. The certificates and their corresponding private keys are used in signing reward transactions.

³33,900 × 10¹⁵ hash function calculations per second. A hash function is the integral part of proof-of-work computation in Bitcoin (Box 2).

⁴19 × 10¹⁵ floating-point operations per second. A floating-point operation is any arithmetic calculations on real numbers. Compared to integer arithmetic in hash functions, floating-point operations are more compute-intensive.

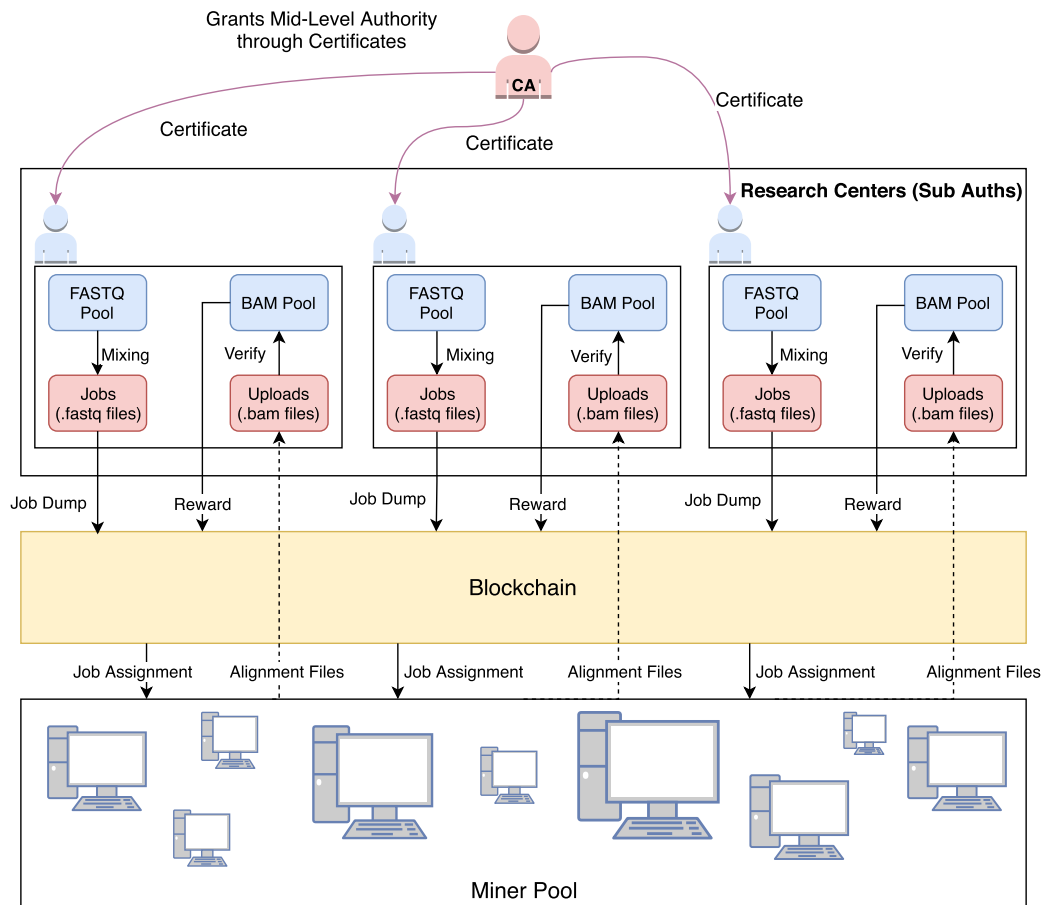


Figure 2. Three-layered structure of Coinami. Root authority issues certificates to research centers, which enables them to distribute HTS jobs to miners. When an HTS job is processed and uploaded successfully, the miner is rewarded with a coinbase transaction that is signed by the subauthority. Before issuing the reward, the subauthority checks whether the alignments are correct using the map location, reference segment, and the string edit fields in the BAM record (i.e., CIGAR and MD); therefore, the generation of the BAM file serves as proof-of-work. These transactions are included in the underlying blockchain. This way, every reward, every transaction is made public so that anyone can inspect the system for a suspicious activity. In this scheme, root authority must be trusted by all parties.

Although Coinami utilizes blockchain mainly as an incentivization scheme, further benefits of open consensus are exploited in the job distribution model. The main authority grants middle authorities the right to publish work tasks in the network. However, job assignment, which is an extension of the naive load balancing approach in grid computing, is distributed among the blockchain users as an integrated smart contract. The extension helps to prevent spam of malicious workers by adding an initial deposit amount to get into the worker pool. This idea also implies that Coinami remains applicable in other blockchains that support smart contracts.

There is still much room for improvement of the Coinami platform, such as difficulty level adjustments similar to that of Bitcoin's, a more general framework to include other applications in omics, and engineering enhancements for optimizing computation load and data transfer at the authorities. We also note that the current multicentralized structure of Coinami due to the need for feeding real alignment problems to the system is not ideal from a blockchain perspective, as blockchain is intended for a completely decentralized usage. However, the rewarding structure is similar to the token mechanism in Ethereum. While minting new coins/tokens is centralized, financial transactions remain de-

centralized. This mechanism still respects the freedom of blockchain, and tokens are proven to be widely acceptable (Catalini and Gans 2018).

There have been different attempts at using the blockchain technology with cryptocurrency as an award mechanism for a computation-heavy task after Coinami was proposed, such as Gridcoin (<http://www.gridcoin.us>), Curecoin (<https://www.curecoin.net>), and FoldingCoin (<http://foldingcoin.net>). Briefly, Gridcoin integrates a cryptocurrency reward into BOINC, in which, unfortunately, genomics and bioinformatics projects are not well represented. Curecoin and FoldingCoin are similar endeavors to bring cryptocurrency incentivization to Folding@home project (Beberg et al. 2009), which is not included in BOINC. Although both projects share a similar base structure following the concepts of Bitcoin, such as mining by hash, they add another layer by integrating Folding@home credits as rewards to miners.

We note that although not directly related to the blockchain technology itself, network bandwidth availability and high volume of input/output operations on the server side may pose scalability issues in compute incentivization approaches. Such problems are always inherent in distributed grid computing platforms.

Privacy-aware data sharing

Data sharing has always been a cornerstone of scientific development. No matter which form or shape it takes, accessible data is essential for reproducibility and further analysis. In the open world of science, it is inevitable to share, access, analyze, and learn from different sources of data for a meaningful result. However, in genomics, the data in question is often personal, private, sensitive, and thus should be treated carefully. The US National Institutes of Health (NIH) issued its Genomic Data Sharing (GDS) policy (National Institutes of Health 2014) to regulate the way that projects generate large-scale human genomic data and use it for subsequent research (<https://osp.od.nih.gov/scientific-sharing/policies/>).

It is challenging to comply with GDS policies while maintaining easy data sharing among researchers. Consequently, genome databases have risen in popularity because of the fact that these databases complied with international policies and provided scientists with a useful medium for data sharing purposes. Although they were useful, centralized approaches like cloud services pose a risk due to their nonfederative control mechanism. A potential problem for data delivery centers was recently revealed in the field of climate science after the 2016 US elections (Dennis 2016). Climate scientists worldwide were concerned that decades of research on climate change could vanish due to confounding political views. This concern prompted them to mirror the data in government controlled centers to independent servers. Similar concerns may rise in shared genomics data, and future political and legal actions may make it impossible to access valuable data even if they were initially made public. Furthermore, most genome databases host varying types of data but enforce the same, most strict policy that is required among all types. Overly firm policies usually end up deterring the end users due to unnecessary diplomatic and bureaucratic steps involved. A decentralized approach in which each owner has complete control over their data—where it is stored, who can access, when it is updated—may be the best way of sharing scientific data. This can be achieved by integrating the consensus model of blockchain into current solutions to decentralized data storage and analysis.

There are currently a few similar proposals to help solve this potential problem for academia. The first one is the Cancer Gene Trust (CGT) being developed by the Global Alliance for Genomics and Health (GA4GH) Consortium (<http://www.cancergenetrust.org>), and the second is the CrypDist project (<https://github.com/CrypDist>). Both projects have similar properties, where summary data such as somatic cancer variation data are kept and distributed in a blockchain. Genomic privacy is achieved through information hiding by sharing only somatic variation and hiding the germline variation. CGT further introduces the concept of “data stewards” who are responsible for inserting only public data to the blockchain and concealing all identifiable information of the patients. CrypDist, on the other hand, also includes mechanisms to share large underlying data (such as BAM files or full-genome VCF files). Note that inserting terabytes or petabytes of additional data to the blockchain is infeasible; instead, CrypDist only keeps links to the large files. Additionally, CrypDist proposes to make use of content delivery networks (CDN) to store and backup these large files to prevent their loss. If the security of these files is of concern, it is possible to further encrypt the files.

Both CGT and CrypDist also aim to encourage data generators to share their valuable data in the system, while providing free and easy access to all researchers. These two aims seem to be in conflict

since a research institute or a hospital may only use available data and never contribute to the system. Resolving this conflict remains an open problem.

Similar to the approaches outlined above, recently announced Gene-chain (<https://www.encryptgen.com/gene-chain/>) and Zenome (Kulemin et al. 2017) offer solutions for genomic data distribution, however with a focus on commercial use of genomic data. In both systems, users freely upload their information to the blockchain. Gene-chain also allows institutions to use the system to collaborate with other researchers, but for a licensing fee. Any research center or company that is interested with any user's data reaches a financial agreement with the data owner (i.e., users), and the data owner grants temporary or permanent access to their data. The financial aspects of the Gene-chain system are not yet fully explained. Furthermore, keeping the entire genome data in the blockchain is infeasible, and how this problem is exactly solved is currently missing from Gene-chain documentation. On the other hand, Zenome is fully described in a white paper (Kulemin et al. 2017), and it uses Ethereum smart contracts to enable data sharing and computation within the same system (Fig. 1). Briefly, users register their data to the Zenome system, and the full data set is kept in a distributed framework, similar to CrypDist. Computational nodes in the system are similar to the Coinami design, and they perform the bioinformatic analyses (i.e., mapping, variation calling, annotation, etc.) as necessary. Here, both computational and storage nodes earn rewards, called “Zenome DNA tokens” (ZNA). The users can buy computational and storage services from the relevant nodes, or they can sell the rights to their data to interested entities such as pharmaceutical companies. Both Gene-chain and Zenome have mechanisms to ensure that the ownership of the data belongs to the users.

Nebula Genomics (<https://www.nebulagenomics.io/>) integrates all aspects of genomic data analysis as an attempt to reenvision genomic data marketplace by building Nebula Network on top of a blockchain. First, Nebula Genomics defines the contemporary genomic data market as a bazaar where the buyers might be sequencing facilities, drug design companies, and healthcare organizations. Individuals pay to obtain information about their genetic variants and possible disease dispositions and get paid by permitting their private data to be shared by third parties in the process. Different from other genetic testing companies that may share their clients' genomic data after providing service discounts, Nebula Network proposes a different market design that prioritizes decisions and privacy of users with the help of Intel's Software Guard Extensions (SGX) and homomorphic encryption on top of blockchain technology. (We omit the details of Intel SGX in this paper; however, we provide a brief overview of homomorphic encryption in Box 1.) As a first step, Nebula Network erases the necessity of personal genomics companies to act as brokers between the data buyers and providers and connects them directly. The economics of the entire system is facilitated by “Nebula Tokens,” which are in fact Ethereum smart contracts. Sequencing facilities agree to accept Nebula tokens to sequence a client's genome, and further management of the data is completely left to the users. They can sell (or rent) their data to other stakeholders in the system such as drug design companies, or researchers, for Nebula Tokens in a secure way using secure compute nodes which are based on homomorphic encryption.

InterPlanetary File System (IPFS) (J Benet, unpubl.) is a distributed file system that aims to bring the web to its decentralized roots. IPFS is essentially a very large control repository. Files are accessible by human-readable addresses to users who have access

rights. It also offers solutions similar to CDN, such as location-based availability of data. Additionally, it is capable of storing sensitive files with encryption and has an access management on top of data distribution. However, rather than being powered by an existing blockchain, IPFS hosts its own blockchain called “FileCoin” (Protocol Labs, unpubl.) to incentivize storage nodes. FileCoin utilizes a proof-of-space algorithm to reward its users. The users are rewarded with FileCoins when they lend storage space. Contrary to most proposals in the blockchain field, both of these projects are currently in use while still being under heavy development. We believe that there is a potential for eliminating data mediators in the genomics field and for connecting data providers and analyzers directly by using ideas similar to IPFS and FileCoin. More recently, another approach was proposed to handle privacy and data liquidity using blockchain (Neisse et al. 2017).

Other possible uses of blockchain

We have described several possible uses of blockchains in distributed computation (e.g., Coinami, FoldingCoin) and data storage and distribution (e.g., CGT, Zenome, Nebula); however, there may be other use cases, not necessarily limited to the life sciences (Table 1). For example, blockchains can be used for online voting such as to agree on proposals within standardization teams. Blockchains can also be used to prove identity and ownership, which may help with a decentralized version of the Open Researcher and Contributor ID (ORCID), or with keeping track of and crediting data ownership.

Another use case may be the Decentralized Autonomous Organizations (DAO), which are defined as virtual and distributed organizations that are governed by a set of rules and contracts represented as Ethereum Smart Contracts (Chohan 2017). Briefly, DAOs are virtual establishments (companies, or otherwise) that make use of the blockchain technology to keep distributed, incorruptible digital ledgers that keep track of financial transactions, agreements, and other sets of rules as required (Vigna and Casey 2016). Although DAOs were first conceptualized as virtual financial organizations, the DAO design may also be used to govern large organizations (e.g., GA4GH, ELIXIR) and international projects (e.g., TCGA, ICGC).

Blockchain: hope or hype?

It is necessary to admit that although blockchain structure was proposed a decade ago, there have been only a very few successful projects like Bitcoin and Ethereum. Additionally, software bugs and hack attempts are still preventing developers from building upon Ethereum (Nikolic et al. 2018). The current widely used blockchain projects are all fundamentally cryptocurrencies. This brings us to question whether blockchain projects function without a monetary base. As we discussed earlier, a token of value must exist to motivate volunteers for proof-of-work or other alternatives.

On the other hand, the technology is still in its infancy. Early adopters were few, and it required a long time for Bitcoin to get traction. In the last year, the field attracted researchers to solve the obstacles that were between being a financial toy and a global currency. Diverse set of cryptocurrencies experimenting with different block sizes, proof concepts, and fault tolerance to achieve better usability on high demand is likely to intensify the fight for market share, which is always a good sign of potential progress.

Final thoughts

Blockchain is a new and exciting technology that might be used to help solve some of the problems we encounter in genomics. There are already several blockchain-based solutions in the field of economics, for example BitPay (<https://bitpay.com>), a middle ground for international money transfer, and OpenBazaar (<https://www.openbazaar.org>), a decentralized marketplace without any fees. However, like almost all new methods, blockchain is not yet mature, and there is still room for further development, especially to ensure cryptographic security.

Furthermore, as we have outlined above, the blockchain technology suffered from negative publicity. Several wallet softwares were hacked, and Bitcoin and other cryptocurrencies were used for illegal activities therefore circumventing the typical checks law enforcement agencies perform to prevent and prosecute criminal action. Also, the very quick rise, followed by a deep crash in Bitcoin’s market value, put a stigma on the blockchain technology as a whole. Technical issues regarding cryptographic security, network bandwidth, proof-of-work or proof-of-stake paradigms, and others can be solved in time. However, for blockchain to be used widely in genomics, it should be well understood that blockchain is not just a cryptocurrency but the underlying technology, and that it might help solve some problems, but it is not a magic spell that can be applied to any computational issues in genomics.

Another advancement that we need for full utilization of the blockchain’s decentralized architecture is decentralization of data itself. As we have mentioned earlier, genomic data needs to be introduced into Coinami, Zenome, and Nebula Genomics systems. Currently only sequencing centers can achieve this, which imposes some level of centralization to these frameworks. This is not ideal because decentralization is the main motivation behind blockchain use. Therefore, to realize the full potential of blockchain for HTS data processing, sequencing should be decentralized. One possible realization of decentralized sequencing might be commoditization of portable sequencers, such as those based on nanopores. However, this remains a hypothetical proposal at this time.

In this Perspective, we tried to speculate on how blockchain can be an integral part of solving several problems in genomics. There are already a few projects in this line of research, which are themselves in their infancy. Many other use cases likely exist for blockchain in scientific computing, data distribution, federated clouds, collaborative work, genome privacy, and others. What could be done, and how, using blockchains remains to be explored.

Competing interest statement

The authors of this manuscript have an ongoing research project, Coinami, that is explained in this Perspective. Additionally, CrypDist was initially proposed as an undergraduate senior design project at the Bilkent University, and C.A. served as project supervisor. However, none of the authors currently own in any cryptocurrencies including Bitcoin, Ethereum, or others, and there are currently no plans to commercialize (i.e., through initial coin offering) either Coinami or CrypDist.

Acknowledgments

We thank R. Durbin, D. Haussler, B. Paten, G. Ratsch, M. Haussler, J. Mattison, R. Chikhi, Ç.T. Öztürk, A. Gökkaya, and B. İcen for

helpful discussions in using blockchain technology in genomics. We also thank N. Lack and M. Somel for their feedback in potential uses of the blockchain as a reward mechanism. This work is partially funded by an EMBO grant (IG-2521) to C.A.

References

- Akcora CG, Gel YR, Kantarcioglu M. 2017. Blockchain: a graph primer. arXiv:1708.08749 [cs.CY].
- Akutsu T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl Math* **104**: 45–62.
- Ayday E, Raisaro JL, Hubaux JP, Rougemont J. 2013. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society*, pp. 95–106, Berlin, Germany.
- Beberg AL, Ensign DL, Jayachandran G, Khaliq S, Pande VS. 2009. Folding@home: lessons from eight years of volunteer distributed computing. In *Proceedings of the 2009 IEEE International Symposium on Parallel and Distributed Processing*, pp. 1–8, Rome, Italy.
- BigchainDB GmbH. 2017. A BigchainDB Primer. <https://www.bigchaindb.com/whitepaper/bigchaindb-primer.pdf>.
- Bonneau J, Miller A, Clark J, Narayanan A, Kroll JA, Felten EW. 2015. SoK: research perspectives and challenges for bitcoin and cryptocurrencies. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pp. 104–121, San Jose, CA.
- Buterin V. 2014. A next-generation smart contract and decentralized application platform. <https://github.com/ethereum/wiki/wiki/White-Paper>.
- Catalini C, Gans JS. 2018. Initial coin offerings and the value of crypto-tokens. National Bureau of Economic Research Working Paper No. 24418, doi: 10.3386/w24418.
- Chohan UW. 2017. The Decentralized Autonomous Organization and governance issues. <https://ssrn.com/abstract=3082055>.
- Croman K, Decker C, Eyal I, Gencer AE, Juels A, Kosba A, Miller A, Saxena P, Shi E, Siler EG, et al. 2016. On scaling decentralized blockchains. In *Proceedings of the International Conference on Financial Cryptography and Data Security*, pp. 106–125. Springer, Berlin, Heidelberg.
- Crosby M, Pattanayak P, Verma S, Kalyanaraman V. 2016. Blockchain technology: beyond Bitcoin. *Appl Innov* **2**: 6–10.
- Dennis B. 2016. “Scientists are frantically copying U.S. climate data, fearing it might vanish under Trump.” *Washington Post*, December 13, 2016.
- Diffie W, Hellman M. 1976. New directions in cryptography. *IEEE Trans Inf Theory* **22**: 644–654.
- Digiconomist 2017. Bitcoin energy consumption index. <https://digiconomist.net/bitcoin-energy-consumption>. Accessed 20 December 2017.
- Flicek P. 2009. The need for speed. *Genome Biol* **10**: 212.
- Gentry C. 2009. “A fully homomorphic encryption scheme.” PhD thesis, Stanford University, Stanford, CA.
- Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351.
- Hart WE, Istrail S. 1997. Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. *J Comput Biol* **4**: 1–22.
- Ileri AM, Ozercan HI, Gundogdu A, Senol AK, Ozkaya MY, Alkan C. 2016. Coinami: a cryptocurrency with DNA sequence alignment as proof-of-work. arXiv:1602.03031 [cs.CE].
- Kulemin N, Popov S, Gorbachev A. 2017. The Zenome Project: blockchain-based genomic ecosystem. <https://zenome.io/download/whitepaper.pdf>.
- McCorry P, Shahandashti SF, Hao F. 2016. Refund attacks on Bitcoin’s Payment Protocol. In *International Conference on Financial Cryptography and Data Security*, pp. 581–599. Springer, Berlin, Heidelberg.
- Metzker ML. 2010. Sequencing technologies—the next generation. *Nat Rev Genet* **11**: 31–46.
- Miraz MH, Ali M. 2018. Applications of blockchain technology beyond cryptocurrency. arXiv:1801.03528 [cs.CR].
- Nakamoto S. 2008. Bitcoin: a peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>.
- National Institutes of Health. 2014. NIH genomic data sharing policy. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html>.
- Neisse R, Steri XJ, Nai-Fovino I. 2017. A blockchain-based approach for data accountability and provenance tracking. arXiv:1706.04507 [cs.CR].
- Nikolic I, Kolluri A, Sergey I, Saxena P, Hobor A. 2018. Finding the greedy, prodigal, and suicidal contracts at scale. arXiv:1802.06038 [cs.CR].
- Raval S. 2016. *Decentralized applications: harnessing Bitcoin’s blockchain technology*. O’Reilly Media.
- Sapirshstein A, Sompolinsky Y, Zohar A. 2016. Optimal selfish mining strategies in bitcoin. In *International Conference on Financial Cryptography and Data Security*, pp. 515–532. Springer, Berlin, Heidelberg.
- Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. 2011. The real cost of sequencing: higher than you think! *Genome Biol* **12**: 125.
- Scherer M. 2017. “Performance and scalability of blockchain networks and smart contracts.” MSc thesis, Umeå University, Umeå, Sweden.
- Spasovski J, Eklund P. 2017. Proof of stake blockchain: performance and scalability for groupware communications. In *The 9th International Conference on Management of Digital EcoSystems. MEDES’17*, Bangkok, Thailand.
- Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big Data: astronomical or genomics? *PLoS Biol* **13**: e1002195.
- Tapscott D, Tapscott A. 2016. Blockchain revolution: how the technology behind Bitcoin is changing money, business, and the world. Portfolio.
- Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.
- Vigna P, Casey MJ. 2016. *The age of cryptocurrency: how Bitcoin and the blockchain are challenging the global economic order*. Macmillan Publishing, London.
- Wang S, Jiang X, Tang H, Wang X, Bu D, Carey K, Dyke SO, Fox D, Jiang C, Lauter K, et al. 2017. A community effort to protect genomic data sharing, collaboration and outsourcing. *NPJ Genom Med* **2**: 33.
- Witte JH. 2016. The Blockchain: a gentle four page introduction. arXiv:1612.06244 [q-fin.GN].
- Yasuda M, Shimoyama T, Kogure J, Yokoyama K, Koshihara T. 2013. Secure pattern matching using somewhat homomorphic encryption. In *Proceedings of the 2013 ACM Workshop on Cloud Computing Security Workshop*, pp. 65–76. Berlin, Germany.



Realizing the potential of blockchain technologies in genomics

Halil Ibrahim Ozercan, Atalay Mert Ileri, Erman Ayday, et al.

Genome Res. published online August 3, 2018

Access the most recent version at doi:[10.1101/gr.207464.116](https://doi.org/10.1101/gr.207464.116)

P<P Published online August 3, 2018 in advance of the print journal.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>
