BLOCKCHAIN™ IN HEALTHCARE TODAY

Accelerating Genomic Data Generation and Facilitating Genomic Data Access Using Decentralization, Privacy-Preserving Technologies and Equitable Compensation

Dennis Grishin,[1,2,3] Kamal Obbad,[1] Preston Estep,[1,3] Kevin Quinn,[1]
Sarah Wait Zaranek,[3] Alexander Wait Zaranek,[1,3] Ward Vandewege,[1,3]
Tom Clegg,[3] Nico César,[3] Mirza Cifric,[1,3] George Church[1,2,3]

Authors
[1]Nebula Genomics, Inc., San Francisco, USA; [2]Department of Genetics, Harvard Medical School, Boston, USA; [3]Veritas Genetics, Inc, Danvers, USA.

Corresponding Author
Dennis Grishin, Nebula Genomics Inc., 73 Sumner Street, #401, San Francisco, CA 94103, USA; dgrishin@g.harvard.edu

Keywords: Arvados, Blockchain, Data Privacy, Data Sharing, DNA Sequencing, Genomics, Homomorphic Encryption, Nebula

Category: Use Cases/Pilots/Methodologies

In the years since the first human genome was sequenced at a cost of over $3 billion, technological advancements have driven the price below $1,000, making personal genome sequencing affordable to many people. Personal genome sequencing has the potential to enable better disease prevention, more accurate diagnoses, and personalized therapies. Furthermore, sharing genomic data with researchers promises identification of the causes of many diseases and the development of new therapies. However, sequencing costs, data privacy concerns, regulatory restrictions, and technical challenges impede the growth of genomic data and hinder data sharing.

In this article, we propose that these challenges can be addressed by combining decentralized system design, privacy-preserving technologies, and an equitable compensation model in a platform that vests control over data with individual owners; ensures transparency and privacy; facilitates regulatory compliance; minimizes expensive data transfers; and shifts the sequencing costs from consumers, patients, and biobanks to researchers in industry and

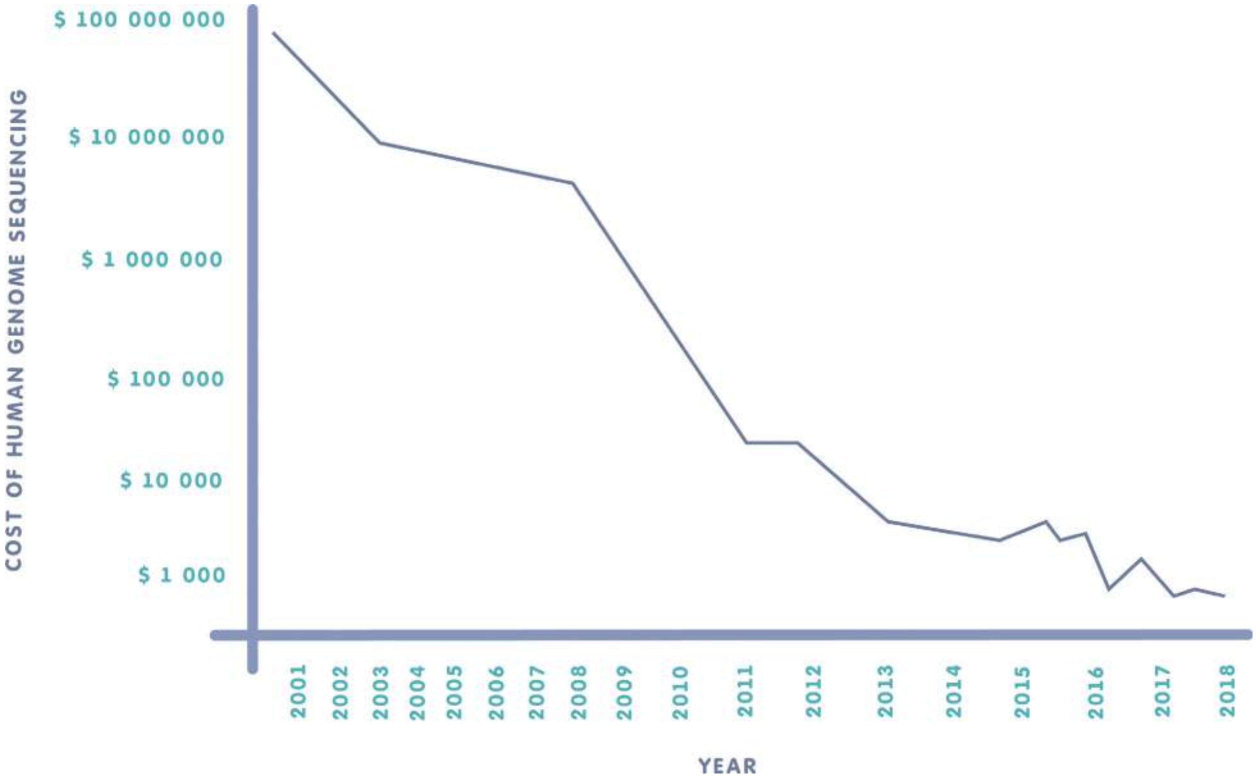

Figure 1—Human genome sequencing cost, 2001–2017.

academia. We exemplify this by describing the implementation of Nebula, a distributed genomic data generation, sharing, and analysis platform.

The Human Genome Project has sequenced and assembled the first human reference genome at a cost of over $3 billion.[1] Since then, development of next-generation sequencing technology has resulted in exponentially decreasing sequencing cost (Figure 1).[2] Today, the sequencing of a whole human genome costs less than $1,000. This price is projected to drop to $100 in the next few years.[3] The exponentially decreasing DNA sequencing costs have made personal genome sequencing affordable to patients as well as healthy individuals.

Personal genome sequencing is becoming more common as prices decline, but most genetic tests to date have been performed using DNA hybridization microarrays. These tests are referred to as genotyping and they assess the presence or absence of genetic variants associated with certain traits. For a cost less than $100, genotyping typically reads out only ~0.02% of the human genome, at predefined positions, often missing health-relevant genetic variants that must be reported. In addition, variant identification at a small number of positions does not allow discovery of novel variants, including those that cause disease; the majority of these variants are distributed throughout the genome and remain undiscovered.[4] This limits the usefulness of genotyping data to researchers.

OPPORTUNITIES

As genomic sequencing becomes more affordable, it opens up opportunities for individuals as well as researchers in academia and industry.

Personal genome sequencing can support data-driven decision-making for health-related issues. Studies estimated that ~2% of people carry genetic variants that cause or predispose them to a wide variety of diseases at various levels of severity, the majority of which can be preventable or treatable.[5] In addition, every parent carries, on average, approximately five genetic variants that might cause diseases in offspring if the other parent carries the same variant.[6] The presence of certain genetic variants also has been associated with adverse effects for ~7% of Food and Drug Administration-approved drugs.[5] Personal genome sequencing can also help healthy individuals make better lifestyle choices. For example, genetic variants have been shown to cause sensitivities to certain nutrients[7–9] and to increase risks of sports-related injuries.[10–12] In the future, advancement in understanding human genetics will make personal genome sequencing more insightful, while correcting pathological genetic variants will become possible as more and more gene therapies enter clinical trials.[13]

Researchers study genomic data sets to identify genetic variants that cause diseases. This enables the research and development of therapies targeting disease-associated genes with increasing specificity. Genomics-guided therapeutic discovery has been applied successfully to many types of cancers, rare genetic diseases, and, increasingly, common complex diseases.[14] Furthermore, genomics-guided patient cohort recruiting can reduce the failure rate of clinical trials by enriching for likely responders and reducing reducing adverse reactions. This approach to clinical trials promises to reduce surging drug development costs and lead to more drugs reaching the market and benefiting patients.[15]

These opportunities are recognized by the biopharma industry. For example, the leading personal genomics company 23andMe received $60 million from Genentech[16] and $300 million from GlaxoSmithKline[17] for access to genotyping data collected from its customers. Other biopharma companies have launched their own sequencing projects. AstraZeneca announced it would sequence 2 million human genomes,[18] and Regeneron is leading a $100 million consortium to sequence approximately 500,000 samples collected by the UK Biobank.[19]

CHALLENGES

Multiple obstacles hinder the realization of opportunities offered by personal genomics. Many people are deterred by the costs of personal genome sequencing, as well as concerns over genomic data privacy. Research is hampered by the resultant scarcity of genomic data and is further compounded by difficulties with respect to data access.

In 2018, the number of genotyped people surpassed 10 million and is expected to grow to more than 100 million by 2021.[20] This growth is driven by a combination of factors, notably consumer interest in ancestry analysis coupled with a decrease in genotyping costs below $100.[21] In contrast, consumer interest in whole genome sequencing has grown slowly due to a significantly higher cost. A recent survey revealed that only ~3% of people are willing to pay >$1,000 for whole genome sequencing.[22] For the majority of consumers, whose primary interest in the area can best be described as nonmedical "infotainment," the benefits of sequencing over genotyping do not justify the significantly higher cost.

At the same time, the surge in popularity of genetic testing, forensic utilization of genetic databases,[23] and the purchase of genetic data by biopharma companies[24] have increased consumer and media attention to genetic data privacy. Studies show that privacy concerns are legitimate, as data sharing policies of many personal genomics companies do not fulfill transparency guidelines with regard to the confidentiality or sharing of customer genetic data.[25] These developments are likely to exacerbate reported privacy concerns over genetic data[26,27] and deter personal genomic sequencing.

For researchers, low adoption of personal genome sequencing has resulted in low availability of genomic data. According to estimates, only ~500 thousand human genomes had been sequenced by 2017.[3] This is detrimental for research because very large genomic data sets are necessary to find links between genetic variants and traits, such as disease predispositions. Finding such links is difficult because most traits are the product of complex interactions of many genetic variants, while the effects of individual genetic variants are, on average, very small.[28] Low diversity of genomic data sets further compounds the search for links between genetics and disease.[29]

The scarcity of genomic data is exacerbated by difficulty in data access due to fragmentation of genomic data across proprietary data silos.[30] Data sharing is further hindered by the large size of genomic data, which impedes data transfer over networks.[31] In addition to logistic and technical challenges, data access is often complicated by restrictive government regulations that hinder data sharing.[32] Low availability of genomic data combined with data silos also results in high prices, making it unaffordable to many researchers.

PREVIOUS WORK

Solutions to the challenges outlined above have been proposed previously. Federated data storage systems have been implemented to facilitate genomic data sharing, privacy-preserving computing has been utilized to protect genomic data privacy, and different compensation models have been explored to incentivize genomic data sharing.

Genomic Data Sharing

The GA4GH Beacon Project[33] and i2b2 SHRINE[34] are two of the most advanced systems for biomedical data sharing. Both are networks that enable participating institutions to connect their genomic (and clinical) databases and process queries about the presence of genetic variants and traits, including medical conditions. This federated model minimizes expensive data transfers and enables institutions to retain control of their data. This addresses privacy, regulatory, and technical challenges that are associated with centralized storage and transfers of genomic data.

However, there are limitations. First, functionality is currently limited to simple queries. Orchestrated, distributed computations required for data processing and analysis are currently not supported. Second, participation is limited to academic research institutions and hospitals. There are no patient- or consumer-focused portals that would enable individuals to easily contribute their personal genomic data. Third, decentralized governance and compensation mechanisms have not been implemented.

Genomic Data Protection

Distributed genomic data storage and computing can help protect genomic data privacy. However, data owners cannot always maintain in-house servers and therefore they often must outsource data storage and computing to third parties, such as cloud service providers. To protect the privacy of genomic data that are shared with untrusted third parties, encryption-based privacy-preserving techniques have been adopted for genomics. These techniques enable third parties to execute computations and return results without having access to plaintext genomic data.

Privacy-preserving techniques have been applied previously to distributed medical and genomic databases. For example, MedCo integrates with the i2b2 SHRINE framework and uses a homomorphic data encryption scheme to enable outsourcing of genomic data storage and query execution to untrusted third parties.[35] Another example is the Secure Multi Party Query Language framework that implements similar functionality and privacy guarantees using secure multiparty computations.[36] Data can also be protected using trusted hardware. An example is the PRINCESS framework that executes computations on genomic data inside protected memory regions of Intel microprocessors.[37]

Compensation Models

Over the past few years, personal genomics companies have explored different models to compensate individuals for contributing their personal genomic data to research studies. In 2016, Genos offered to help its customers sell their genomic data to researchers.[38] A similar model that uses a cryptocurrency instead of fiat money was adapted by EncrypGen in 2017.[39] Most recently, LunaDNA announced that it would compensate genomic data contributors with company stock.[40] These models are similar in that individuals who want to participate must already own their personal genomic data, or choose to purchase genetic testing because of the prospect they will be rewarded later for sharing the data.

PERSONAL GENOMICS 2.0

The traditional model for genomic data generation and sharing that has been adopted by most personal genomics companies contributes to the challenges described in the previous sections. This model requires consumers to pay for genetic testing and result interpretation, while personal genomics companies often take ownership of the generated genomic data and sell it to biopharma companies (Figure 2). This model requires consumers to carry the costs and relinquish ownership and control of their genomic data, which discourages genetic testing. In addition, this model promotes genomic data fragmentation across private data silos, which hampers data access and increases data prices.

We propose to combine and extend previous work on genomic data sharing networks, privacy-preserving technologies, and compensation models to create a new model for personal genomics that may overcome these challenges (Figure 3).

First, the functionality of genomic data sharing networks must be extended beyond simple queries. This requires a network that can be integrated with a full-fledged bioinformatics platform that supports genomic data processing and analysis. Implementing this functionality would bundle fragmented genomic data and make it available for analysis on a single network, thereby facilitating data access for researchers.

Second, the data sharing network must expand beyond research institutions and must be accessible to individuals who want to share their personal genomic data. However, the resulting network decentralization will necessitate a more democratic governance model. This potentially can be achieved by integrating blockchain technology, which holds the promise of enabling decentralized, self-governing networks.

Third, the privacy of genomic data must be protected. Data access control on the blockchain

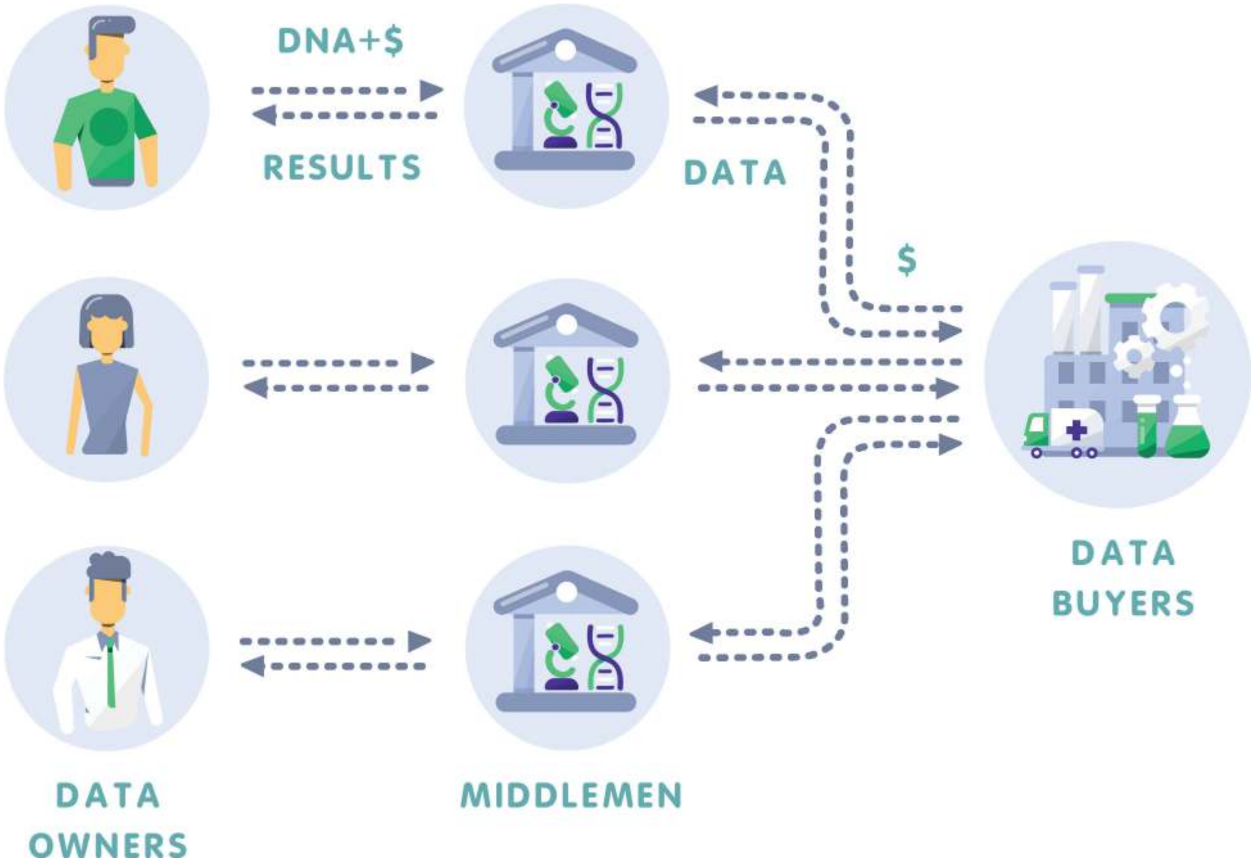

Figure 2. The traditional model for genomic data generation and sharing.

can ensure transparent consent management, while privacy-preserving technologies can help protect shared genomic data. Together with the distributed computing model that "brings algorithms to the data," these technologies can enable network participants to retain ownership and control of their genomic data, thereby reducing privacy concerns and incentivizing data sharing.

Fourth, genome sequencing and data sharing also must be incentivized by implementing subsidy and compensation mechanisms. The decentralized data sharing model can facilitate this, as it enables researchers to connect directly with individuals with traits of interest, subsidize their genome sequencing costs, and compensate them for data sharing. Elimination of middlemen also may result in a reduction in genomic data prices and thus empower more researchers to access large genomic data sets.

DESIGN CONSIDERATIONS

To implement a system as outlined in the previous section, one must integrate a bioinformatics platform that supports distributed data storage and computing with a suitable blockchain framework, as well as with techniques for privacy-preserving computing. Here, we review and evaluate existing options.

Bioinformatics Platforms

Bioinformatics platforms have been developed to facilitate organization of genomic data; to enable parallelized, high-performance computing with support for complex dependencies; and to allow

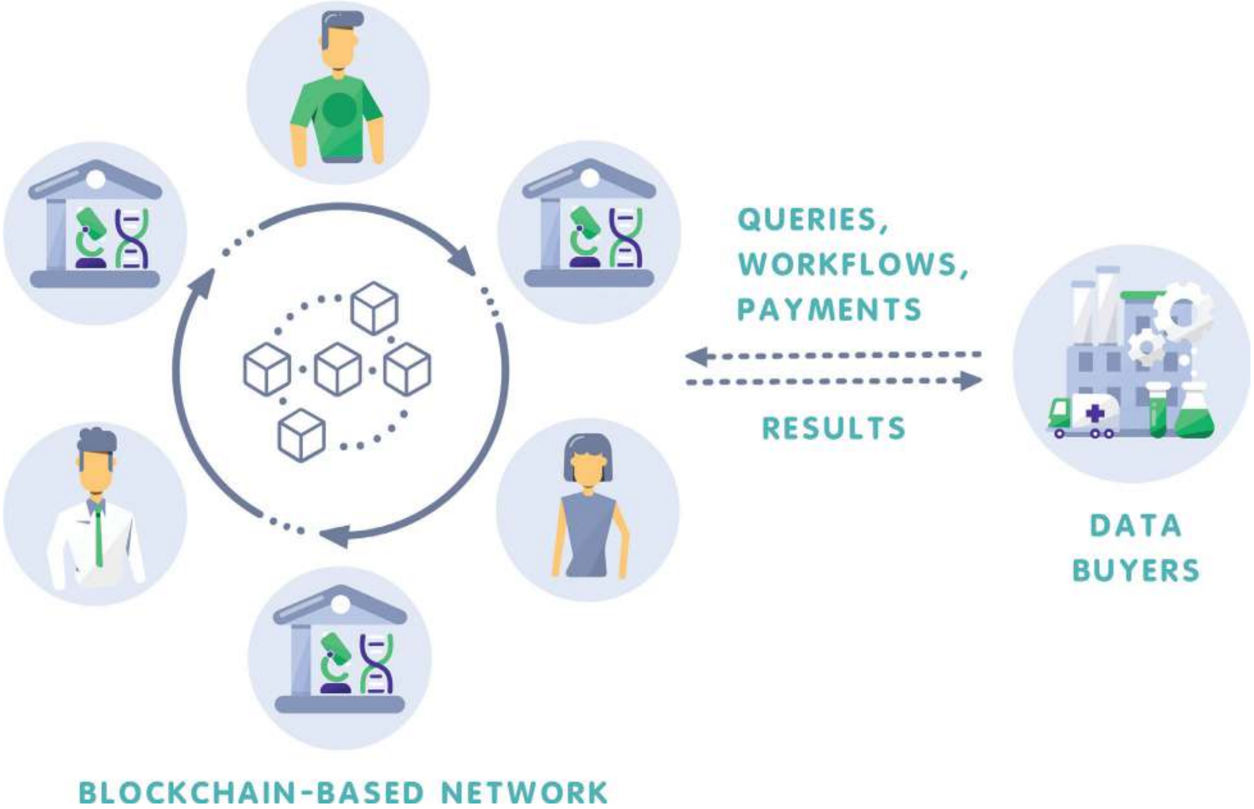

Figure 3—Alternative model for personal genomics that may overcome challenges.

a modular pipeline design that is flexible and ensures reproducible results.[41] Table 1 shows a comparison of popular bioinformatics platforms.

The development of bioinformatics platforms has been driven by exponentially growing genomic data and marked by adaption of multiple computing trends. Storage and processing of genomic data has moved from local servers to remote clouds. This has enabled scalable data storage and computing and facilitated access sharing to genomic data sets. To scale beyond single clouds, efforts are being made to create federated cloud environments that could enable distributed data storage and computing.[48,49] Furthermore, the growth of genomic data and development of new bioinformatics tools that must be integrated into workflows are driving the development of standardized workflow description languages, containerization of computing environments, and utilization of standardized application programming interfaces (APIs).

Based on these considerations, Arvados and DNAstack appear as suitable choices for the proposed genomic data sharing platform. Both platforms have an API-focused architecture and data sharing functionality. DNAstack integrates with the GA4GH Beacon Network, while Arvados supports platform-agnostic, federated cloud environments and has an open-source codebase.

Blockchain Frameworks

Blockchain technology has three use cases in the proposed system. First, the need to provide transparent consent management can be addressed by the ability of blockchains to store data access permissions on an

Table 1. Comparison of bioinformatics platforms

Criteria	Arvados[42,43]	DNAstack[44]	Seven Bridges[45]	DNAnexus[46]	Galaxy[47]
Hardware	Federated clouds and servers	Google Cloud with Beacon Network integration	Clouds	Clouds	Local servers
Pipeline design	API-based; web GUI	API-based; web GUI	Web GUI	Web GUI	Web GUI
Containers	Yes	Yes	Yes	Yes	Yes
Workflow language	CWL	WDL	CWL	Custom	Custom
Open source	Yes	No	No	No	Yes
Platform launch year	2013	2014	2012	2010	2005

API: application programming interface; CWL: Common Workflow Language; GUI: graphical (rather than textual) user interface; WDL: Workflow Description Language.

immutable public ledger. Second, blockchains can enable implementation of decentralized systems governed by network participants. Third, an immutable ledger can facilitate verification of the integrity of decentrally stored data.

Based on these use cases, one can create a set of requirements that a suitable blockchain framework must fulfill. First, consent management requires that the identity of researchers who request to access data are known to data owners. To this end, network access must be limited to data buyers whose identity has been verified. Therefore, consent management requires a blockchain that supports permissioned access.

Second, a large, decentralized data marketplace requires smart contract functionality and high transaction throughput. Private blockchains can achieve higher transaction throughputs than public blockchains because the ability to write transactions to the blockchain is limited to a group of permissioned validator nodes. However, this makes private blockchains more centralized and less dependable.

Based on these requirements, permissioned blockchains frameworks such as Exonum and Hyperledger Fabric appear most suitable (Table 2). Hyperledger Fabric has been more widely adopted, but Exonum offers transparency and security that is comparable to public blockchains.

First, Exonum-based blockchains offer public read access but restrict write access to selected validator nodes. By making read access to the blockchain public, transaction audit does not rely on trusted parties. Exonum transactions are verified in real time by all nodes. Thus, all network participants are able to audit the blockchain state collectively.

Second, Exonum supports anchoring of transaction logs in the Bitcoin blockchain. Hashes of the Exonum blockchain state are periodically written to the Bitcoin blockchain, so even if all permissioned Exonum nodes collude, the transaction history cannot be falsified unless the attacker succeeds in compromising the Bitcoin blockchain as well.

Third, Exonum uses a byzantine fault-tolerant (BFT) consensus algorithm that protects against

Table 2. Comparison of blockchain frameworks

Criteria	Exonum[50]	Hyperledger Fabric[51]	Ethereum[52]
Read access	Public	Private	Public
Write access	Private	Private	Public
Consensus	Byzantine fault-tolerant (BFT)	Fault-tolerant (FT)	Proof of work (PoW)
Transactions per second (TPS)	~3,000	~3,000	~15
Smart contracts	Yes (Rust, Java)	Yes (Go, Java)	Yes (Solidity)
Light clients	Yes	No	Yes
Public blockchain anchoring	Yes	No	NA
Open source	Yes	Yes	Yes

NA: not applicable.

malicious behavior of permissioned nodes. In contrast, Hyperledger and other private blockchains rely on less computationally intensive fault-tolerant (FT) consensus algorithms that protect against node breakdown but not malicious behavior. Exonum offers both BFT consensus and high transaction throughput because it is written in Rust, one of the fastest programming languages. Furthermore, Rust offers memory safety which eliminates many vulnerabilities that are commonly exploited by hackers.

Privacy-Preserving Technologies

Table 3 shows a comparison of privacy-preserving technologies that all have been applied to secure genomic data.[53] Fully homomorphic encryption and secure multiparty computations enable computations on encrypted data that generate encrypted results. These encrypted results, when decrypted, correspond to the results of the same computation on plaintext data. However, fully homomorphic encryption is very slow and typically suffers from very large ciphertext expansion. The limitation of secure multiparty computation protocols is that they require transfers of very large data amounts during the computation. It is possible, however, to improve the performance of fully homomorphic encryption and secure multiparty computations significantly if they are optimized for specific use cases. Practical performance levels have been demonstrated for queries on genomic data[54,55] and genome-wide association studies (GWAS).[56]

Alternative technologies have drawbacks of their own. Intel Software Guard Extensions technology is a hardware-assisted approach that protects data privacy by executing computations inside private memory regions. It offers good performance but has been affected by vulnerabilities that can compromise data privacy.[57] Differential privacy methods protect data privacy by introducing randomness. However, obfuscation of computation results can complicate interpretation of studies.[53]

NEBULA

In this section, concepts and design considerations outlined in the previous sections are illustrated by describing the technical implementation of Nebula—a decentralized genomic data generation, sharing, and analysis

Table 3. Comparison of privacy-preserving technologies

Criteria	Fully Homomorphic Encryption	Secure Multiparty Computations	Intel Software Guard Extensions	Differential Privacy
Principle	Computations (additions AND multiplications) on ciphertexts	Distributed computations on ciphertexts	Computations inside private memory regions	Introduction of randomness to data/results of computations
Computation time	Very slow	Slow	Fast	Fast
Memory usage	Very high	High	Low	Very low
Communication cost	High	Very high	Low	Low
Specific limitations	None	None	Vulnerabilities have been discovered; requires Intel CPUs	Noise makes interpretation of results more difficult

CPU: central processing unit.

platform. Nebula integrates the Arvados[42,43] bioinformatics platform (github.com/curoverse/arvados) with the Exonum[50] blockchain framework (github.com/exonum) and a fully homomorphic data encryption scheme (Figure 4).

Arvados has two core services: Keep and Crunch. Keep is a distributed content-addressable storage system that enables scalable storage of genomic big data, high throughput data access, and efficient data management. Crunch is a workflow management engine that enables flexible creation and parallelized execution of data analysis pipelines and generation of reproducible results. Arvados implements a distributed data storage and computing model that minimizes required data transfers. This helps address big data challenges, regulatory restrictions, and data privacy risks.

Utilization of a homomorphic data encryption scheme enables implementation of privacy-preserving queries on genomic data. The intention is to preserve data privacy by enabling investigators to query the whole database and discover their data of interest, without compromising the privacy of the queried data. In the future, it should be possible to extend the application of privacy-preserving technologies to GWAS and other computations.

The Nebula blockchain is an Exonum-based blockchain through which the Nebula network will be governed, consent will be documented, and the data will be secured. Exonum-based blockchains have three types of nodes: auditors, light clients, and validators. Auditors are full nodes that maintain a copy of the entire blockchain content and can generate transactions. Light clients also can generate transactions, but they replicate only information that is relevant to them instead of the whole blockchain content. Validators are permissioned nodes that verify transactions received from auditors and light clients and write new blocks to the blockchain. While the current implementation of Nebula

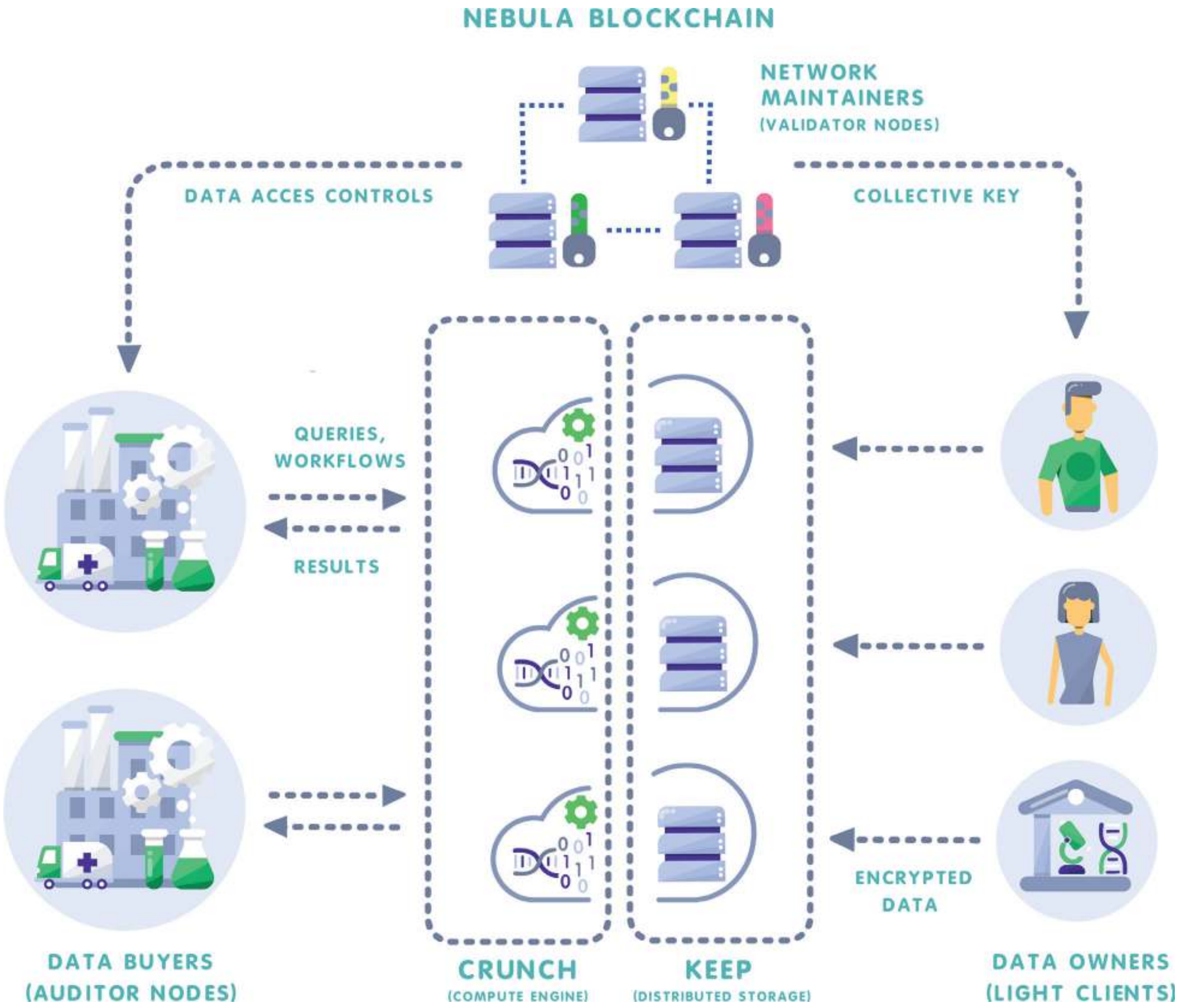
NEBULA BLOCKCHAIN
NETWORK MAINTAINERS (VALIDATOR NODES)
DATA ACCES CONTROLS
COLLECTIVE KEY
QUERIES, WORKFLOWS
RESULTS
ENCRYPTED DATA
DATA BUYERS (AUDITOR NODES)
CRUNCH (COMPUTE ENGINE)
KEEP (DISTRIBUTED STORAGE)
DATA OWNERS (LIGHT CLIENTS)

Figure 4—Overview of the Nebula platform.

uses the Exonum framework, other permissioned blockchains, in particular Hyperledger Fabric, can be used as well.

The Nebula network has four types of participants: data owners, network maintainers, data buyers, and storage and compute providers.

- *Data owners* can be private individuals or institutions. They will store encrypted genomic data in public or private clouds that are part of the Keep storage system. They will be able to control access to their data and receive payments to their wallets by operating light clients on the Nebula blockchain.
- *Network maintainers* are organizations that operate validator nodes on the Nebula blockchain. Validator nodes will collectively control data access by managing encrypted key shares, verifying transactions, and keeping track of data stored in Keep and computations executed by Crunch.
- *Data buyers* are researchers who wish to obtain access to genomic data. They will be operating auditor nodes to keep a local copy of

the metadata, which they will use to locate data stored in Keep, verify data integrity, and keep track of access permissions. Data buyers will be able to query homomorphically encrypted data, utilize smart contracts to acquire data access permissions from data owners, and use Crunch to run analysis pipelines.

- *Storage and compute providers* are data owners that operate private clouds, or third parties that offer storage and computing services (e.g., Google, Amazon, and Microsoft). They will form a federated cloud environment that hosts the Keep storage system and Crunch-managed containers within which computations are executed.

The development of Nebula is ongoing. Some parts of the platform, in particular, Arvados, have been fully implemented over the past few years and are already being deployed by various organizations. Other parts of Nebula, in particular, the homomorphic encryption schemes, are a relatively recent addition and are not yet fully integrated. A report on the progress of our work was published in a white paper.[58] Here we describe the implementation of Nebula in greater detail but also revise some previously made design choices.

Data Generation

Genomic data

Personal genome sequencing cost is a significant factor in preventing more widespread consumer adoption. Therefore, a key consideration in the design of the Nebula platform was to provide a mechanism to shift sequencing costs from data owners (e.g., consumers and biobanks) to data buyers (e.g., pharma and biotech companies). This is being implemented by enabling data buyers to query the Nebula database, identify data sets of interest, and pay the sequencing costs to generate and access genomic data (Figure 5). To this end, the Nebula platform enables a data buyer to create a smart contract that specifies the blockchain addresses of data owners previously identified in a query and send cryptocurrency tokens to that smart contract. The data owners are notified that a buyer has offered to pay their sequencing costs. If a data owner accepts the offer by executing the smart contract, the deposited tokens are sent to a sequencing provider. Next, the data owner receives a saliva collection kit and submits a saliva sample to the sequencing facility. The sample is sequenced, and the genomic data are deposited on a Keep server specified by the data owner. Data hashes, along with blockchain addresses of all data owners and buyers, are written to the blockchain. The data buyer who paid the sequencing costs is permitted to access and analyze the data. The data owner receives interpretations of his genomic data and is able to share data access with additional data buyers.

Phenotypic data

Information about medical conditions and other traits is referred to as phenotypic data. These data are generated primarily through survey questions. The platform utilizes a phenotyping toolkit that maps plain-language survey responses to clinical descriptions in Human Phenotype Ontology (HPO)[59] format. Survey data can be verified using two approaches. First, comparing the incidence of medical conditions in the general population to the incidence observed in the platform’s data set will enable identification of survey results that deviate from the expected results. Second, survey data can be verified by referencing Electronic Health Records (EHRs) imported through the Fast Healthcare Interoperability Resources (FHIR) API.

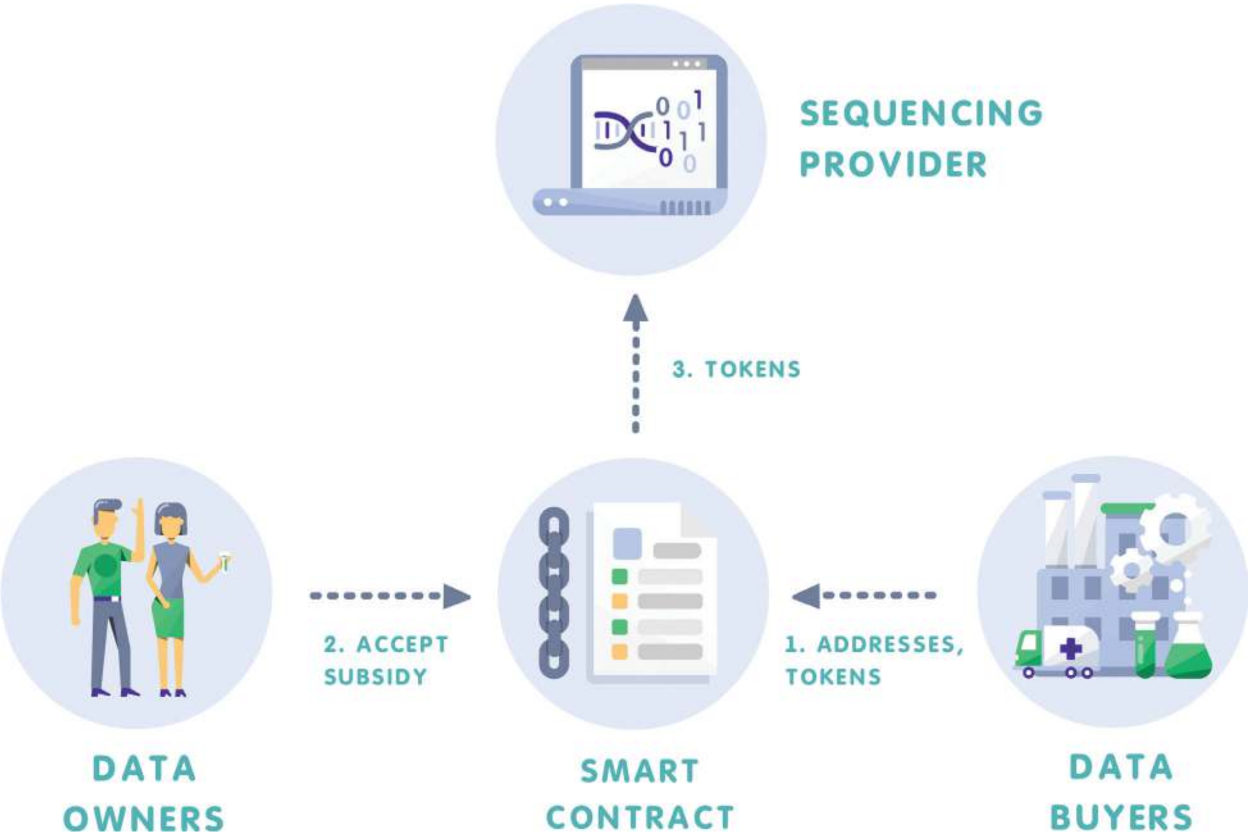

Figure 5—Genome sequencing subsidy payment on the blockchain.

Data Encryption

Privacy of genomic and phenotypic data are protected through client-side encryption by data owners and encryption key management by blockchain validator nodes. To enable data buyers to discover data prior to purchasing data access, the platform implements a lattice-based fully homomorphic encryption scheme. To this end, blockchain validator nodes generate public–private key pairs and construct a single collective public key (Figure 4). Data owners encrypt their survey responses and genetic variant lists with the collective public key and upload them to a Keep server. The homomorphic encryption scheme protects data privacy by enabling data buyers to execute Structured Query Language (SQL)-like queries on the homomorphically encrypted data. Files that contain raw sequencing data and are not used for queries are Advanced Encryption Standard (AES) encrypted. The AES keys are encrypted with validator public keys and bundled with the encrypted data.

Data Storage

Data

Genomic data are stored in Keep, a distributed content-addressable storage system that retrieves files based on their content. Addresses of files are generated through cryptographic digest of their content. Keep combines content-addressing with the distributed storage architecture of the Google File System.[60] Keep splits encrypted files into 64-megabyte blocks and stores them in an underlying object store or file system (Figure 6). The content addresses of the blocks are stored on the blockchain and are used to find data locations and check data integrity.

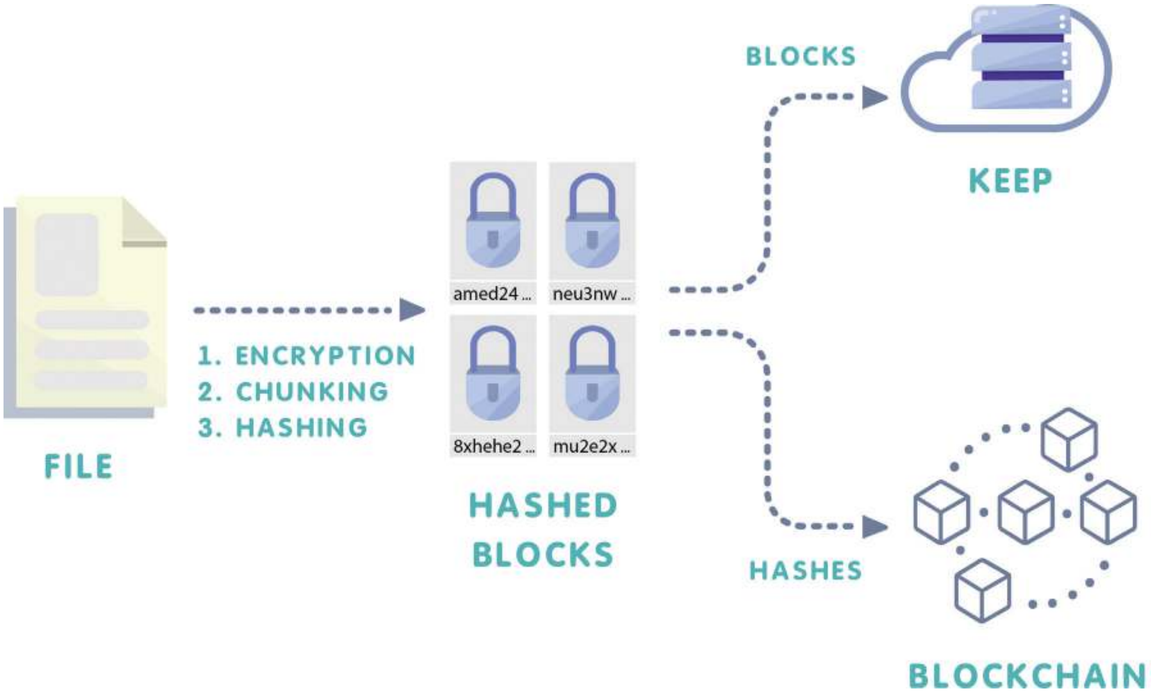

Figure 6—Data blocks are stored in Keep. Block hashes are stored on the blockchain.

Keep is designed for storing genomic and other types of biomedical big data. First, its content-addressing offers high-speed storage and retrieval by eliminating an indexing service, a potential bottleneck and point of failure, and enabling direct connections between the storage and compute subsystems. Second, content-addressing works well for data written to disk once and read many times, a characteristic of genomic data, as it does not change over time but is accessed frequently. Third, fixed-size data blocks allow scalable distributed storage of big data, and content-addressing enables easy file verification, which is particularly important for distributed databases.

Keep is designed to be a distributed, hybrid storage system. Data owners can choose to store their data in clouds such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure, or on private bare-metal servers. Decentralized file storage solutions such as InterPlanetary File System (IPFS), Sia, and Storj can potentially be supported if computing on stored data becomes possible. Data owners can register new, personal cloud instances or store their encrypted data in shared clouds. Based on phenotypic information, data sets that are likely to be analyzed together are stored in physical proximity, which minimizes slow and expensive data transfers.

As sequencing data are processed, different file formats are generated and stored in Keep. Typically, Keep stores FASTQ files that contain raw sequencing data (~200 gigabytes/genome), Binary Alignment Map files that store aligned sequencing reads (~100 gigabytes/genome), and Variant Call Format files that store genetic variants (~200 megabytes/genome). Additionally, Nebula uses the Compact Genome Format (CGF) to generate compact genomic data summaries. Genomes in the CGF format are represented by pointers referencing sequences in a tile library (Figure 7). CGF offers a consistent, standardized representation of genomic data that makes

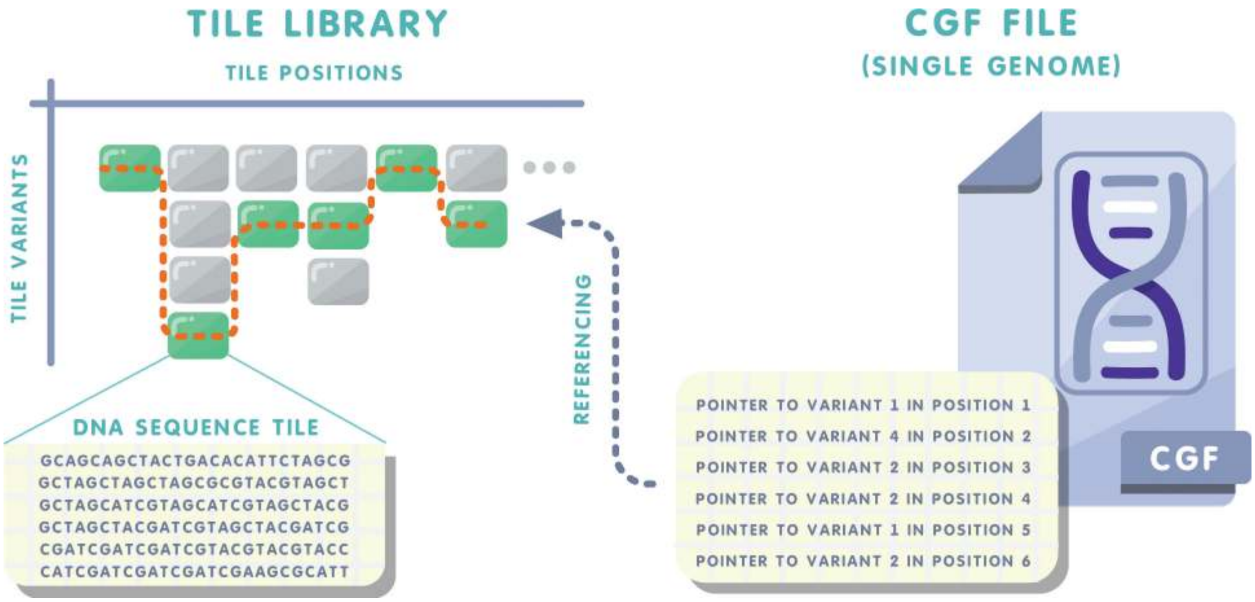
TILE LIBRARY
TILE POSITIONS
TILE VARIANTS
DNA SEQUENCE TILE
GCAGCAGCTACTGACACATTCTAGCG
GCTAGCTAGCTAGCGCGTACGTAGCT
GCTAGCATCGTAGCATCGTAGCTACG
GCTAGCTACGATCGTAGCTACGATCG
CGATCGATCGATCGTACGTACGTACC
CATCGATCGATCGATCGAAGCGCATT
REFERENCING
CGF FILE
(SINGLE GENOME)
POINTER TO VARIANT 1 IN POSITION 1
POINTER TO VARIANT 4 IN POSITION 2
POINTER TO VARIANT 2 IN POSITION 3
POINTER TO VARIANT 2 IN POSITION 4
POINTER TO VARIANT 1 IN POSITION 5
POINTER TO VARIANT 2 IN POSITION 6
CGF

Figure 7—Simplified representation of a tile library and a Compact Genome Format (CGF) file. The rectangles represent tile variants at different positions and the dotted line illustrates the tile composition of specific genome.

different types of sequencing and genotyping data interoperable. The CGF representation is also very space efficient (~30 megabytes/genome), which facilitates file transfers, and enables fast queries and efficient analysis.[60]

Tabular phenotypic data generated through surveys and imports of EHRs are stored in physical proximity with associated genomic data. In contrast to static genomic data, phenotypic information is much more dynamic and smaller than genomic data. This makes utilization of the Google File System and content addressability unsuitable. Therefore, phenotypic data files are stored as Not only SQL documents.

Metadata

To organize data stored in Keep, Nebula stores metadata on the blockchain in a key-value store. When new data are added to Keep or existing data are modified, blockchain transactions are generated. Validator nodes verify these transactions, add new blocks to the blockchain, and update the key-value store. Storage of metadata on an immutable ledger helps secure the integrity of the decentralized Nebula database. To this end, multiple column families are implemented:

- *Data ownership* is registered by assigning each block content address the blockchain address of the data owner who added the block to Keep.
- *Data locations* are described by assigning each block content address the Uniform Resource Locator (URL) of a Keep server.
- *Data integrity* is verified by re-hashing data blocks and comparing their hashes with content addresses that are stored on the blockchain.
- *Data buyer identities,* including names and institutional affiliations, are verified, linked to blockchain addresses, and stored on the blockchain.
- *Access permissions* to the Nebula platform and data stored in Keep also are managed on the blockchain.

Data Discovery

Utilization of fully homomorphic encryption is intended to address the privacy barrier to data sharing. It enables data owners to make their data available for discovery without privacy risks, while at the same time allowing data buyers to explore the database before purchasing data access to perform analyses.

To this end, data buyers will construct a SQL-like query and encrypt it with the collective public key that has been constructed by validator nodes and used to encrypt phenotypic information and genetic variant lists. The encrypted query is executed on homomorphically encrypted data and an encrypted result is generated. The query result is re-encrypted by the validator nodes under a public key provided by the data buyer and shared with data buyer who can now decrypt it with its private key. A query can return the number of data owners that matched the specified criteria, as well as their blockchain addresses. This enables data buyers to connect with data owners to pay sequencing costs or to purchase access to existing genomic data (Figure 8).

Data Analysis

The Arvados container and pipeline management engine, Crunch, executes computations on data stored in Keep. Crunch implements a distributed computing model whereby workflows, and not the genomic data, are moved between cloud instances whenever possible. Highly distributed genomic data processing is possible because many intensive bioinformatics computations, such as alignment and variant calling, are performed on single genomes and are easily parallelizable. To this end, Crunch executes tasks inside Docker containers that are created physically close to the data locations in Keep and distributes computations between many processing units.

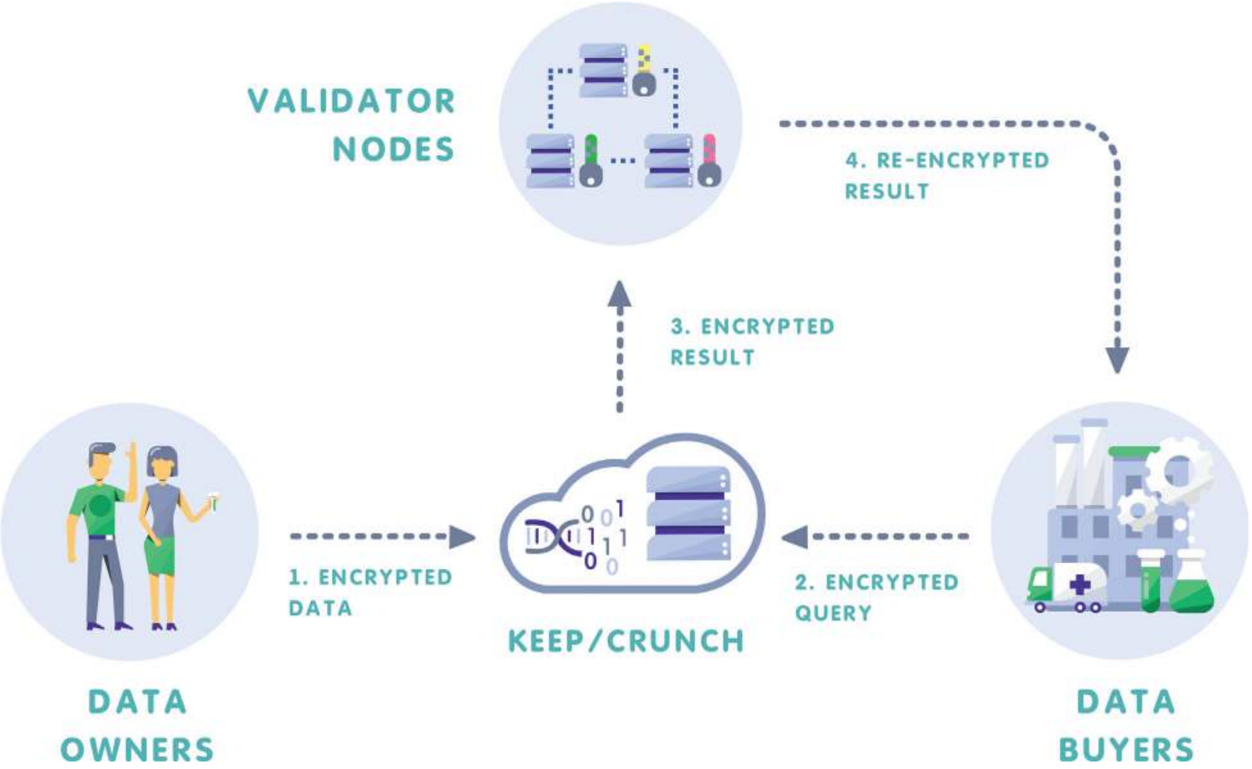

Figure 8—Secure data discovery through queries on homomorphically encrypted data.

Crunch ensures result reproducibility through standardization of computing workflows using Common Workflow Language (CWL),[61] which enables connection of open-source and proprietary bioinformatics software into workflow pipelines that are flexible, portable, and scalable. Crunch can access CWL pipelines stored in public or private Git repositories such as GitHub.

CWL can be used to implement end-to-end bioinformatics analysis pipelines. Typically, CWL pipelines include common, computationally intensive "secondary analysis" tasks, such as alignment of sequencing reads to a reference genome and variant calling. However, "tertiary analysis" tasks, which often involve computing on genomic data sets rather than single genomes and are less standardized, also can be incorporated into CWL pipelines. Typical examples are statistical tests that are used in GWAS to identify correlations between genetic variants and phenotypes. For such tertiary analysis tasks, Nebula uses Lightning,[62] a system for high-performance, in-memory computations on genomic data in the CGF. Lightning integrates into CWL pipelines and enables fast queries and execution of complex machine learning tasks on large genomic data sets.

CWL pipelines can also be used to analyze and interpret personal genomic (and phenotypic) data. First, users can build their own custom pipelines to analyze their personal data and also share pipelines among each other using public Git repositories. Second, developers can build and monetize genomic apps. To this end, CWL pipelines can be stored in private repositories, and access by Crunch may require a smart contract-mediated token payment to the pipeline developer. The approach of bringing apps to the data facilitates protection of personal information.

Security

Homomorphic encryption can enable privacy-preserving queries for data discovery. However, most computations that are necessary for typical genomic data analysis workflows do not achieve practical runtimes when executed on homomorphically encrypted data. Therefore, other security mechanisms must be utilized.

Platform access control

To protect data owners and their data, data buyers are required to go through a partially decentralized, three-step permission process. The first step is data buyer authentication. Here, a blockchain validator node will verify a data buyer's personal and institutional identity. Blockchain addresses of verified data buyers will be added to the blockchain metadata store. Data buyers will then be able to connect to Nebula REST API servers and use Crunch to execute pipelines on data stored in Keep. Data buyer authentication will enable data owners to verify data buyer identity before agreeing to share data access. Furthermore, immutable storage of data buyer identities on the blockchain enables identification of data buyers who have violated consent agreements or have bypassed pipeline execution control.

Pipeline execution control

To protect data privacy, the platform design incorporates the ability to define approved bioinformatics tools and CWL workflows. The intent is to prevent data buyers from downloading genomic data or executing any computations that attempt to extract a large amount of information about individual data owners. This approach was chosen because it has the ability to provide a sufficient level of data privacy protection without significantly restricting data buyers.

Data access control

The first task in every CWL pipeline is to get access to the input data (Figure 9). Here, a data buyer executes a smart contract on the blockchain. The inputs are the data buyer's blockchain address and the content addresses of all data blocks of the input files. The data buyer also deposits tokens inside the smart contract and defines a token payout for data access. When a data owner's light client synchronizes with the blockchain, the data owner is notified of the data access request. The data owner can decide about data sharing based on offered payment and identity of the data buyer. The data owner grants data access by executing the smart contract. The blockchain validator nodes then verify the integrity of the requested data stored in Keep by comparing data hashes with the content addresses stored on the blockchain and collectively re-encrypt the data under the data buyer's public key. Finally, the data buyer's access permission is registered on the blockchain and tokens are sent from the smart contract to the data owner's wallet. Crunch can now load decrypted data into a Docker container and begin pipeline execution.

Governance

The Nebula blockchain can be used to enable Nebula network participants to collectively govern the network, in particular, to help maintain data protection. To this end, for example, Token-Curated Registries (TCRs)[63] can be used to conduct elections that determine validator nodes or whitelist data analysis pipelines. TCRs are lists that are curated decentrally by token holders. Importantly, economic incentives drive the token holders to curate the list's contents judiciously. In brief,

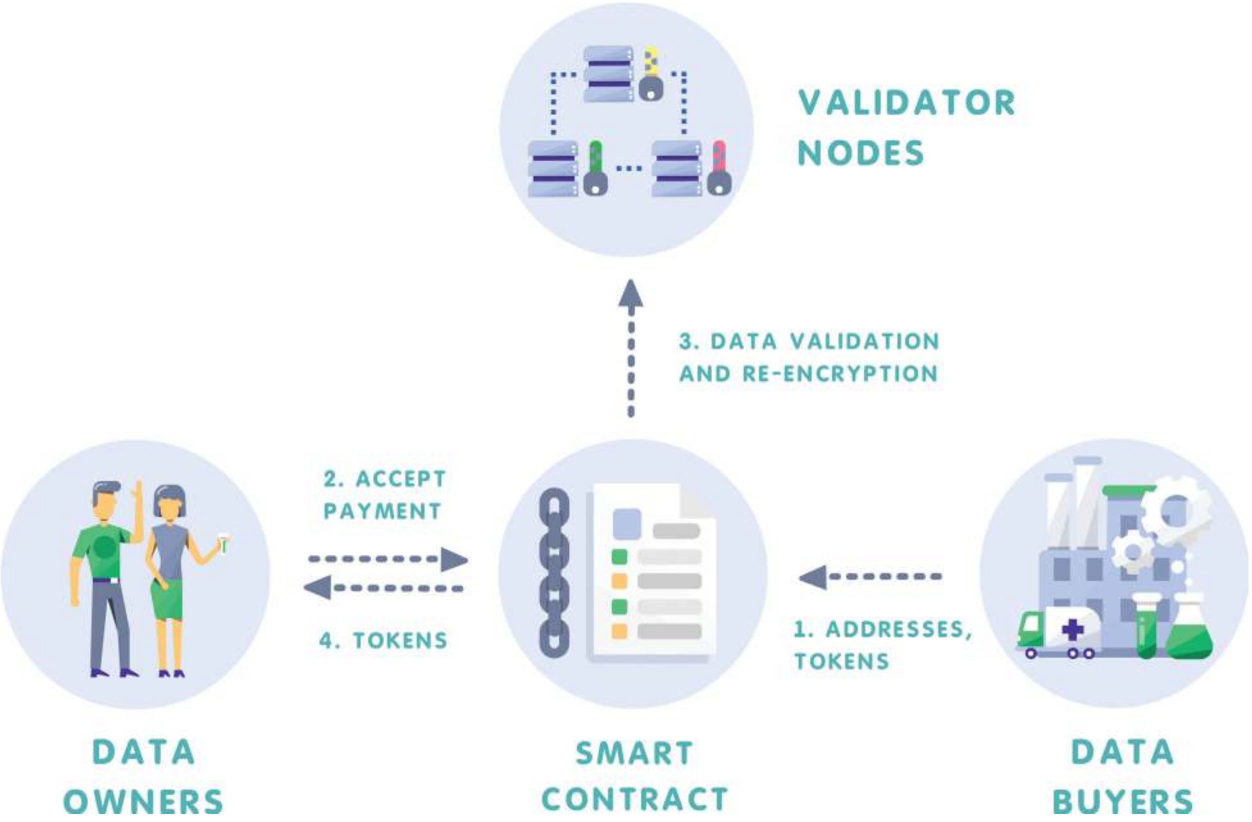

Figure 9—Data access control and data purchases.

network participants can cast votes whereby the weights of votes scale with their token holdings. Since token holders are invested in the network, they are incentivized to maintain its proper operation that ensures data protection.

DISCUSSION

The obstacles that hinder personal genome sequencing and genomic data sharing have a significant impact on the progress of research into disease prevention, drug development, and other crucial aspects of human health. We described one approach to overcoming these obstacles, using a combination of multiple technologies. A number of challenges remain to be addressed regarding data privacy, data validation, data curation, and data economics.

Data privacy protection requires decentralization of data generation and further development of privacy-preserving technologies. Today, genomic data generation is limited to laboratories that own expensive sequencing machines operated by experienced technicians. Centralized genomic data generation leads to data privacy risks that may be averted if sequencing is decentralized. We anticipate that this will become possible soon as new technologies are being developed that would enable compact, affordable, and easy-to-operate sequencing machines.[64] Data privacy protection is also impaired by current limitations of privacy-preserving technologies that do not allow complex computations on large data sets and require extensive optimization for every use case, which hinders effective data analysis. However, practicality of privacy-preserving technologies has been steadily increased over the past few years, and we anticipate continuing progress in the future.

Data validation and curation are another area of challenge. Validation of genomic data requires assistance of the sequencing facilities that have produced the data. If the source of the genomic data is unknown, or the sequencing facility does not cooperate, genomic data cannot be validated. A possible solution to this problem can be a model that compensates personal genomics companies and other genomic data producers for validating data authenticity. Data collected from different sources also must be made interoperable. It is particularly challenging to curate health records and other types of phenotypic data. However, standards such as FHIR are being developed very actively and have already enabled applications that can collect EHRs across different health systems.[65]

The idea of a personal data marketplace is very new and has not yet been implemented at scale. A personal data marketplace is likely to differ from traditional marketplaces in important ways. For example, data supply can be regarded as being unlimited because an individual can share data access with an unlimited number of data buyers. Personal data marketplaces also would be asymmetric, since individuals are likely to be unaware of the value of their personal data and are thus at risk of not being compensated fairly. The novelty of personal genomic data further compounds these challenges and makes market dynamics difficult to predict. We anticipate that future research into economics of data marketplaces will help answer these and other open questions.

Contributions: Dennis Grishin, Kamal Obbad, and Kevin Quinn wrote the article. Dennis Grishin, Kamal Obbad, and George Church developed the ideas described in the article. Dennis Grishin, Kamal Obbad, and Kevin Quinn are leading the development of the Nebula platform. Alexander Wait Zaranek, Ward Vandewege, Tom Clegg, Nico César, Preston Estep, and Mirza Cifric contributed to the development of Arvados. Alexander Wait Zaranek

and Sarah Wait Zaranek developed the Compact Genome Format and Lightning. All authors edited and/or reviewed the final article.

Acknowledgments

The authors would like to thank Armon Rahim and Nathaniel Tucker for their help in reviewing this article.

Conflict of Interest

All authors are employees, advisors, or collaborators of Nebula Genomics.

Funding Statement

Development of the Nebula platform is funded by Nebula Foundation.

REFERENCES

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature.* 2004 Oct 21;431(7011):931–45.
2. Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* [Internet]. [cited 2018 Jan 11]. Available from: https://www.genome.gov/sequencingcostsdata/
3. *Illumina Promises to Sequence Human Genome For $100—But Not Quite Yet*. 2017. [cited 2018 Oct 10]. Available from: https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/#672a5d72386d
4. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012 Sep 7;337(6099):1190–5.
5. Lindor NM, Thibodeau SN, Burke W. Whole-genome sequencing in healthy people. *Mayo Clin Proc*. 2017 Jan;92(1):159–72.
6. Berg JS, Adams M, Nassar N, et al. An informatics approach to analyzing the incidentalome. *Genet Med*. 2013 Jan;15(1):36–44.
7. Yang A, Palmer AA, de Wit H. Genetics of caffeine consumption and responses to caffeine. *Psychopharmacology.* 2010 Aug;211(3):245–57.
8. Mattar R, de Campos Mazo DF, Carrilho FJ. Lactose intolerance: Diagnosis, genetic, and clinical factors. *Clin Exp Gastroenterol*. 2012 Jul 5;5:113–21.
9. Freeman HJ. Risk factors in familial forms of celiac disease. *World J Gastroenterol*. 2010 Apr 21;16(15):1828–31.
10. O'Connell K, Knight H, Ficek K, et al. Interactions between collagen gene variants and risk of anterior cruciate ligament rupture. *EJSS.* 2015;15(4):341–50.
11. Tiziano FD, Palmieri V, Genuardi M, Zeppilli P. The role of genetic testing in the identification of young athletes with inherited primitive cardiac disorders at risk of exercise sudden death. *Front Cardiovasc Med*. 2016 Aug 26;3:28.
12. Bennett ER, Reuter-Rice K, Laskowitz DT. Genetic Influences in Traumatic Brain Injury: Chapter XI. In: Laskowitz D, Grant G, editors. *Translational Research in Traumatic Brain Injury.* Boca Raton, FL: CRC Press/Taylor and Francis Group; 2015.
13. Ginn SL, Amaya AK, Alexander IE, Edelstein M, Abedi MR. Gene therapy clinical trials worldwide to 2017: An update. *J Gene Med*. 2018 May;20(5):e3015.
14. Cardon LR, Harris T. Precision medicine, genomics and drug discovery. *Hum Mol Genet.* 2016 Oct 1;25(R2):R166–72.
15. Rojahn SY. Genomics could blow up the clinical trial. *MIT Technology Review* [Internet]. 2013 Nov 12 [cited 2018 Aug 25]; Available from: https://www.technologyreview.com/s/521496/genomics-could-blow-up-the-clinical-trial/
16. Herper M. Surprise! With $60 Million Genentech Deal, 23andMe Has A Business Plan [Internet]. *Forbes.* 2015 [cited 2017 Oct 1]. Available from: https://www.forbes.com/sites/matthewherper/2015/01/06/surprise-with-60-million-genentech-deal-23andme-has-a-business-plan/

17. Bloomberg. GlaxoSmithKline Is Acquiring a $300 Million Stake in 23andMe [Internet]. *Fortune*. [cited 2018 Aug 25]. Available from: http://fortune.com/2018/07/25/glaxosmithkline-23andme-gsk/
18. Ledford H. AstraZeneca launches project to sequence 2 million genomes. *Nature.* 2016 Apr 28;532(7600):427.
19. Herper M. Drug company consortium to sequence the genes of 500,000 Britons over next two years. Forbes Magazine [Internet]. 2018 Jan 8 [cited 2018 May 27]; Available from: https://www.forbes.com/sites/matthewherper/2018/01/08/drug-company-consortium-to-sequence-the-genes-of-500000-britons-over-next-two-years/
20. Khan R, Mittelman D. Consumer genomics will change your life, whether you get tested or not. *Genome Biol*. 2018 Aug 20;19(1):120.
21. Allyse MA, Robinson DH, Ferber MJ, Sharp RR. Direct-to-Consumer Testing 2.0: Emerging models of direct-to-consumer genetic testing. *Mayo Clin Proc.* 2018 Jan;93(1):113–20.
22. Marshall DA, Gonzalez JM, Johnson FR, et al. What are people willing to pay for whole-genome sequencing information, and who decides what they receive? *Genet Med*. 2016 Dec;18(12):1295–302.
23. Kolata G, Murphy H. The golden state killer is tracked through a thicket of DNA, and experts shudder. *The New York Times* [Internet]. 2018 Apr 27 [cited 2018 Aug 21]; Available from: https://www.nytimes.com/2018/04/27/health/dna-privacy-golden-state-killer-genealogy.html
24. Ducharme J. A major drug company now has access to 23andMe's genetic data. Should you be concerned? *Time* [Internet]. 2018 Jul 26 [cited 2018 Aug 21]; Available from: http://time.com/5349896/23andme-glaxo-smith-kline/
25. Laestadius LI, Rich JR, Auer PL. All your data (effectively) belong to us: Data practices among direct-to-consumer genetic testing firms. *Genet Med.* 2017 May;19(5):513–20.
26. Bloss CS, Ornowski L, Silver E, et al. Consumer perceptions of direct-to-consumer personalized genomic risk assessments. *Genet Med.* 2010 Sep;12(9):556–66.
27. Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al. Public attitudes toward consent and data sharing in Biobank Research: A large multi-site experimental survey in the US. *Am J Hum Genet.* 2017 Mar 2;100(3): 414–27.
28. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: Biology, function, and translation. *Am J Hum Genet*. 2017 Jul 6;101(1):5–22.
29. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016 Oct 13;538(7624):161–4.
30. Lawler M, Maughan T. From Rosalind Franklin to Barack Obama: Data sharing challenges and solutions in genomics and personalised medicine. *New Bioeth.* 2017 Apr;23(1):64–73.
31. Feltus FA, Breen JR 3rd, Deng J, et al. The widening gulf between genomics data generation and consumption: A practical guide to big data transfer technology. *Bioinform Biol Insights.* 2015 Sep 23;9(Suppl 1):9–19.
32. Majumder MA, Cook-Deegan R, McGuire AL. Beyond our borders? Public resistance to global genomic data sharing. *PLoS Biol.* 2016 Nov;14(11):e2000206.
33. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. *Science.* 2016 Jun 10;352(6291):1278–80.
34. Weber GM, Murphy SN, McMurry AJ, et al. The Shared Health Research Information Network (SHRINE): A prototype federated query tool for clinical data repositories. *J Am Med Inform Assoc.* 2009 Sep;16(5):624–30.
35. Raisaro JL, Troncoso-Pastoriza J, Misbach M, et al. MedCo: Enabling secure and privacy-preserving exploration of distributed clinical and genomic

data. *IEEE/ACM Trans Comput Biol Bioinform* [Internet]. 2018 Jul 13; Available from: http://dx.doi.org/10.1109/TCBB.2018.2854776

36. Bater J, Elliott G, Eggen C, Goel S, Kho A, Rogers J. SMCQL: Secure querying for federated databases. *Proceed VLDB Endowment.* 2017 Feb;10(6):673–84.
37. Chen F, Wang S, Jiang X, et al. PRINCESS: Privacy-protecting rare disease International Network Collaboration via encryption through software guard extensions. *Bioinformatics*. 2017 Mar 15;33(6):871–8.
38. Molteni M, Allain R, Chen S, Thompson A, Simon M, Gonzalez R. Genos will sequence your genes—And help you sell them to science. *Wired* [Internet]. 2016 Dec 15 [cited 2018 Oct 6]; Available from: https://www.wired.com/2016/12/genos-will-sequence-genes-help-sell-science/
39. Lin P. Blockchain: The missing link between genomics and privacy? *Forbes* [Internet]. 2017 May 8 [cited 2018 Oct 6]; Available from: https://www.forbes.com/sites/patricklin/2017/05/08/blockchain-the-missing-link-between-genomics-and-privacy/
40. Brown KV. Share your DNA, get shares: Startup files an unusual offering. *Bloomberg News* [Internet]. 2018 Oct 5 [cited 2018 Oct 6]; Available from: https://www.bloomberg.com/news/articles/2018-10-05/illumina-backed-startup-asks-sec-to-let-it-pay-people-for-dna
41. Leipzig J. A review of bioinformatic pipeline frameworks. *Brief Bioinform.* 2017 May 1;18(3):530–6.
42. Arvados Documentation [Internet]. [cited 2018 Oct 10]. Available from: doc.arvados.org
43. Zaranek AW, Clegg T, Vandewege W, Church GM. Free factories: Unified infrastructure for data intensive web services. *Proc USENIX Annu Tech Conf.* 2008 May 1;2008:391–404.
44. DNAstack Documentation [Internet]. [cited 2018 Oct 9]. Available from: https://docs.dnastack.com/java-sdk/
45. Seven Bridges Documentation [Internet]. [cited 2018 Oct 10]. Available from: docs.sevenbridges.com/docs
46. DNAnexus Documentation [Internet]. [cited 2018 Oct 10]. Available from: wiki.dnanexus.com
47. Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010 Aug 25;11(8):R86.
48. Chaterji S, Koo J, Li N, Meyer F, Grama A, Bagchi S. Federation in genomics pipelines: Techniques and challenges. *Brief Bioinform* [Internet]. 2017 Aug 29; [cited 2018 Oct 10]. Available from: http://dx.doi.org/10.1093/bib/bbx102
49. Workflow Execution Service (WES) API [Internet]. Github; [cited 2018 Oct 11]. Available from: https://github.com/ga4gh/workflow-execution-service-schemas
50. Exonum Documentation [Internet]. [cited 2018 Oct 10]. Available from: exonum.com/doc
51. Androulaki E, Barger A, Bortnikov V, et al. Hyperledger fabric: A distributed operating system for permissioned blockchains. In: *Proceedings of the Thirteenth EuroSys Conference*. New York: ACM; 2018. pp. 30:1–30:15. (EuroSys ’18).
52. Wood G. Ethereum: A secure decentralised generalised transaction ledger. 2014. [Internet]. [cited 2018 Oct 10]. Available from: https://ethereum.github.io/yellowpaper/paper.pdf
53. Aziz MMA, Sadat MN, Alhadidi D, et al. Privacy-preserving techniques of genomic data-a survey. *Brief Bioinform* [Internet]. 2017 Nov 7; [cited 2018 Oct 10]. Available from: http://dx.doi.org/10.1093/bib/bbx139
54. Çetin GS, Chen H, Laine K, et al. Private queries on encrypted genomic data. *BMC Med Genomics.* 2017 Jul 26;10(Suppl 2):45.
55. Sousa JS, Lefebvre C, Huang Z, et al. Efficient and secure outsourcing of genomic data storage. *BMC Med Genomics*. 2017 Jul 26;10(Suppl 2):46.
56. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty

computation. *Nat Biotechnol.* 2018 Jul;36(6):547–51.

57. Chen G, Chen S, Xiao Y, et al. *SgxPectre Attacks: Stealing Intel Secrets from SGX Enclaves via Speculative Execution* [Internet]. arXiv [cs.CR]. 2018. [cited 2018 Oct 10]. Available from: http://arxiv.org/abs/1802.09085
58. Grishin D, Obbad K, Estep P, et al. *Nebula—Blockchain-Enabled Genomic Data Sharing and Analysis Platform* [Internet]. [cited 2018 Oct 10]. Available from: https://www.nebula.org/assets/Nebula_Genomics_Whitepaper.pdf
59. Robinson PN, Köhler S, Bauer S, et al. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* 2008 Nov;83(5):610–5.
60. Ghemawat S, Gobioff H, Leung S-T. The Google file system. In: *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles (SOSP '03)*. New York: ACM; 2003. pp. 29–43.
61. Amstutz P, Crusoe M, Tijanić N, et al. *Common Workflow Language, v1.0. Specification, Common Workflow Language working group*. 2016. [Internet]. [cited 2018 Oct 10]. Available from: https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156/2
62. Guthrie S, Connelly A, Amstutz P, et al. Tiling the genome into consistently named subsequences enables precision medicine and machine learning with millions of complex individual datasets [Internet]. *PeerJ PrePrints*; 2015 Oct [cited 2018 Jan 16]. Report no.: e1780. Available from: https://peerj.com/preprints/1426/
63. Goldin M. Token-Curated Registries 1.0—Mike Goldin—Medium [Internet]. Medium. Medium; 2017 [cited 2018 Oct 10]. Available from: https://medium.com/@ilovebagels/token-curated-registries-1-0-61a232f8dac7
64. Erlich Y. A vision for ubiquitous sequencing. Genome Res. 2015 Oct;25(10):1411–6.
65. EHRIntelligence. *Breaking Down How the Apple Health Records EHR Data Viewer Works* [Internet]. EHRIntelligence. 2018 [cited 2018 Oct 10]. Available from: https://ehrintelligence.com/news/breaking-down-how-the-apple-health-records-ehr-data-viewer-works