Blockchain-enabled genomic data sharing and analysis platform

Dennis Grishin

Kamal Obbad

Preston Estep

Mirza Cifric

Yining Zhao

George Church

v4.52

01/25/2018

Decoding the human genome sequence is the most significant undertaking that we have mounted so far in an organized way in all of science. I believe that reading our blueprints, cataloguing our own instruction book, will be judged by history as more significant than even splitting the atom or going to the moon.

Francis S. Collins
Director of the National Institutes of Health (NIH)

TABLE OF CONTENTS

1. Abstract

The first human genome was sequenced in 2001 at a cost of $3 billion. Today, human genome sequencing costs less than $1000, and in a few years the price will drop below $100. Thus, personal genome sequencing will soon be widely adopted as it enables better diagnosis, disease prevention, and personalized therapies. Furthermore, if genomic data is shared with researchers, the causes of many diseases will be identified and new drugs developed. These opportunities are creating a genomic data market worth billions of dollars.

Nebula Genomics seeks to lead this emerging market by understanding and overcoming key obstacles. We will spur genomic data growth by significantly reducing the costs of personal genome sequencing, enhancing genomic data protection, enabling buyers to efficiently acquire genomic data, and addressing the challenges of genomic big data. We will accomplish this through decentralization, cryptography, and utilization of the blockchain.

SEQUENCING COSTS - The Nebula peer-to-peer network will enable data buyers to acquire genomic data directly from data owners without middlemen. This will enable data owners to receive sequencing subsidies from data buyers and profit from sharing their data. Reducing sequencing costs will accelerate growth of genomic data.

DATA PROTECTION - Data owners will privately store their genomic data and control access to it. Shared data will be protected through zero-trust, encryption-based secure computing. Data owners will remain anonymous, while data buyers will be required to be fully transparent about their identity. The Nebula blockchain will immutably store all data transaction records. Addressing data privacy concerns will likewise accelerate growth of genomic data.

DATA ACQUISITION - The Nebula network will aggregate genomic data from individuals and genomic databanks and thereby address the problem of genomic data fragmentation. Direct communication with data owners and a smart survey tool will enable data buyers to collect high quality trait information. Utilization of standardized data formats will facilitate data curation. Smart contracts will automate and thereby accelerate data purchases.

GENOMIC BIG DATA - Genomic big data is projected to outgrow video and text data within the next few years. Through decentralized data storage, flexible utilization of available computing power and efficient file transfers enabled by space-efficient data encoding, the Nebula network will absorb the forthcoming data explosion.

2. Introduction

2.1. DNA, genes and genomes

Deoxyribonucleic acid, or DNA, is a molecule that encodes the blueprint of every living organism. DNA is a chain-like molecule of variable length made of four building blocks, commonly called letters (Figure 1). The four letters of DNA are adenine (A), thymine (T), cytosine (C), and guanine (G). Thus, while computers encode data using the binary system, the code of life uses four symbols.

The total amount of DNA found in a cell is referred to as its genome. Genes are DNA sequences that encode instructions for the production of proteins - molecular machines with a wide variety of functions. The human genome consists of ~ 6.4 billion letters that encode ~ 20,000 protein-coding genes and a much larger number of functional, non-coding elements. The roles of most functional sequences in the human genome are still unknown.

Figure 1. DNA sequence.

2.2. DNA sequencing

Since the discovery of the function and structure of DNA, scientists have been seeking to develop methods to read DNA sequences in order to study and eventually understand the code of life. The sequencing of the first reference human genome was completed in 2003 [1]. The project lasted more than a decade and cost over $3 billion. At this price, certainly no more than a few human genomes would have ever been sequenced. Fortunately, over the past 10 years, a technological quantum leap has taken place. The laboratory of Nebula Genomics co-founder Prof. George Church [2], as well as other laboratories, have developed methods for high throughput, "next generation" DNA sequencing. While "first generation" sequencing machines were designed to read the DNA sequence of a single DNA molecule, current next-generation sequencing machines read billions of molecules in parallel.

As a result of this unprecedented technological advancement, the cost of DNA sequencing has been plummeting at an exponential rate (Figure 2) [3]. In 2017, the cost to produce and interpret a clinical-quality whole genome sequence broke below the key $1000 price barrier. This achievement was accomplished by Veritas Genetics (veritasgenetics.com), which was founded and is managed by members of the Nebula Genomics team. The price for human whole genome sequencing is expected to drop to as little as $100 within the next few years [4]. In the meantime, the sequencing cost can be reduced through targeted sequencing of protein-coding genome regions, whereby non-coding regions are left out. This sequencing method, called exome sequencing, currently costs ~ $300.

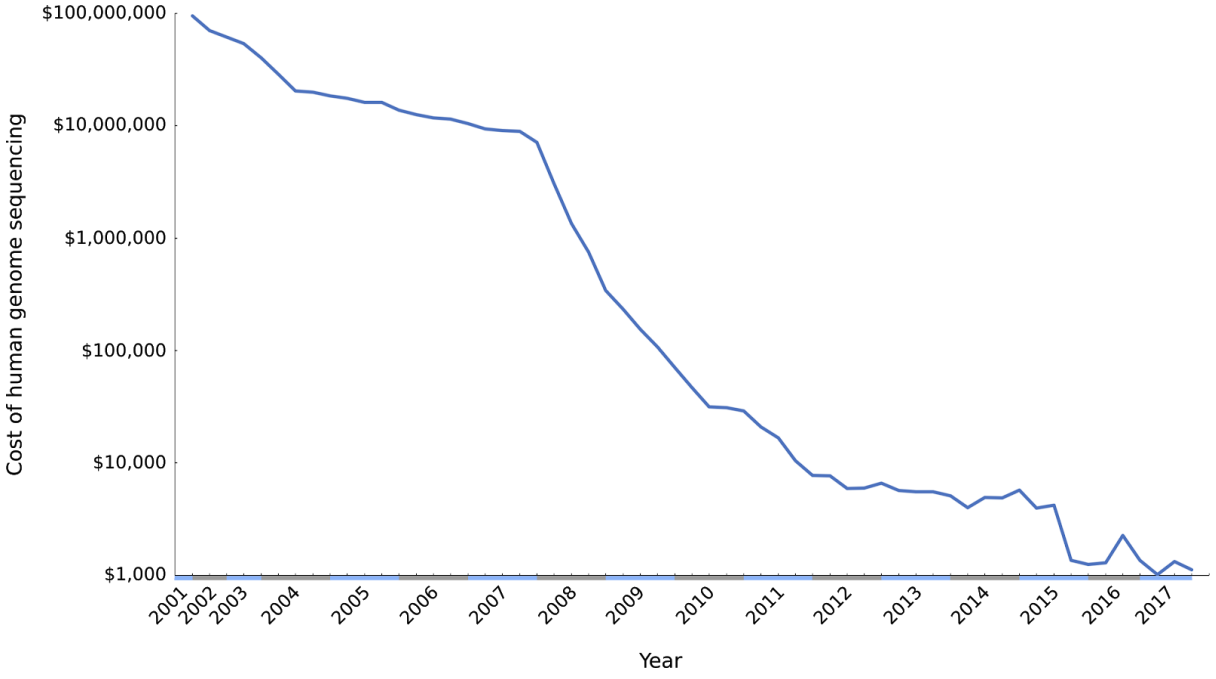

Figure 2. Human genome sequencing cost 2001 - 2017.

2.3. Sequencing vs. genotyping

Today, 23andMe (23andme.com) and Ancestry (ancestry.com) are the two leading personal genomics companies. Both use DNA microarray-based genotyping for their genetic tests. It is an outdated and significantly less powerful alternative to DNA sequencing. Instead of sequencing continuous stretches of DNA, genotyping identifies single letters spaced at approximately regular intervals across the genome. While human whole genome sequencing determines ~ 6.4 billion letters, microarray-based genotyping used by 23andMe and Ancestry identifies letters at only ~ 600.000 positions. Thus it generates small amounts of data that are of limited value to individual data owners and researchers.

Individuals can learn much more about their personal genetic makeup from whole genome sequencing data. Additionally, as researchers learn more about human genetics, interpretations of personal genomic data can be continuously updated and new insights gained. Whole genome sequencing data is also much more useful to researchers than other data types. In particular, whole genome sequencing is the only method for identifying non-coding DNA variants. Since over 90% of DNA regions of clinical importance fall in non-coding regions, whole genome sequencing is likely to become the primary means by which many therapeutic targets will be discovered [5]. In research laboratories, sequencing has already largely replaced microarray-based genotyping and the personal genomics market will follow suit. Thus, companies that rely on inferior genotyping technology will lose their share of the personal genomics market.

3. Opportunities

3.1. Individuals

Any two humans are, genetically, 99.9% identical to each other. However, the 0.1% difference comprises over 4 million genetic variants [6] that are responsible for differences we observe between people, including physical traits, personality and disease predispositions. With personal genome sequencing it is now possible to fully characterize what makes each one of us unique. This information can be used to make optimal choices in regard to all health-related questions including medical treatments, family planning, personal diet, and exercise regimen. We are now entering the age of more precise and personalized medicine and healthcare, with an increasing focus on prevention.

Medical prescriptions - More than 7% of FDA-approved medications are affected by genetic variants that cause some patients to react to the medication adversely [7]. Knowledge of individual pharmacogenes can aid physicians in determining the proper choice and dosage of medication to prescribe to patients. For example, warfarin, a commonly prescribed blood thinning drug, can cause fatal internal bleedings in individuals who carry genetic variants that potentiate its blood thinning effect [8].

Preventive treatments - Approximately 2% of people carry early-onset pathogenic variants in highly "actionable" genes [7]. These are genes associated with a pathology for which treatments exist and are likely to alter the individual's outcome. For example, mutations in *BRCA1* and *BRCA2* genes dramatically increase the risk of breast and ovarian cancer. Hence women who carry these mutations are strongly advised to undergo frequent screenings [9].

Very high percentages of people suffer from less deadly but still problematic genetic variants. Knowledge of such variants allows preventive action to be taken before a disease becomes serious. For example, fatty liver disease is estimated to affect over 80 million people in the U.S. [10]. It can remain undetected for many years, leading to high rates of liver failure and cancer. Genetic variants carried by over half the people of certain populations greatly increase the risk of serious complications of fatty liver disease.

Risks of other diseases can be increased substantially by genetic variants that are even more common. For example, the existence of rare DNA variants that protect against diseases of old age reveal that the vast majority of people carry DNA variants that predispose them to typically treatable common diseases of older age, such as cardiovascular disease, dementias, and diabetes [11].

Family planning - Two prospective parents may want to sequence their genomes in order to understand their risk of having a child with a genetic disorder. Genome sequencing of an adult individual typically detects ~ 5.5 genetic variants that do not affect the individual but may cause a disease in the offspring if disease-associated variants are inherited from both parents [12]. It has been estimated that about 350 million people, or 5% of the world's population, suffer from a

genetic disease or disorder [13]. Most of these conditions are inherited from the parents, and can be detected through whole genome sequencing.

Diet - Genetic variants have been found to influence effectiveness of weight loss strategies [14], protection against dietary carcinogens [15], sensitivities to nutrients such as caffeine [16], lactose [17], and gluten [18], and predispositions to vitamin deficiencies [19–21].

Physical exercise - Genetic variants are also associated with physical performance including endurance and muscle strength [22] as well as risks of sports injuries. For example, the risk of ligament tears is linked to variants in collagen genes [23], physical activity can lead to heart failure in genetically predisposed individuals [24], and head blows in sports like football and boxing significantly increase the risk of degenerative brain diseases if certain genetic variants are present [25].

Gene editing - The identification of genetic variants that contribute to physical traits and disease predispositions is the first step towards genetic engineering that would eliminate diseases and confer desirable traits. The laboratory of Nebula Genomics co-founder Prof. George Church developed CRISPR-Cas9 [26], a revolutionary genome editing tool. Numerous promising target genes have already been identified. For example, deactivation of the myostatin gene could potentially cure degenerative muscle diseases and, in healthy individuals, lead to less body fat, more muscle mass, and a healthier cardiovascular system [27,28]. The introduction of genetic variants that are protective against diseases of old age holds substantial promise to extend human lifespan.

3.2. Industry

Genomic datasets can be used to identify associations between genetic variants and diseases. This enables rational design of drugs that will modulate selected genes and produce therapeutic effects. Genomic data-driven drug design also enables personalized medicine - drugs made for patients with a particular genetic makeup. These opportunities are creating a multi-billion dollar genomic data market. For example, in 2012 Amgen acquired the genomics company deCODE for $415M [29]. deCODE has sequenced the genomes of 2,636 Icelanders to discover genetic variations associated with disease [30]. In 2015 23andMe received $60M from Genentech for access to its genetic databank [31]. Due to lack of human genome sequencing data, biotech companies are also launching their own sequencing projects. In 2016 AstraZeneca announced the launch of a project to sequence 2 million human genomes [32], and in 2018 Regeneron announced the creation of a $50M consortium to sequence the exomes of 500,000 samples from the UK Biobank [33]. Furthermore, in an effort to enable personalized healthcare, hospitals have begun to offer personal genome sequencing to patients. In 2018 the Mayo Clinic started a preventive healthcare initiative and has entered a partnership with Veritas Genetics to make whole genome sequencing broadly available to apparently healthy individuals [34].

4. Challenges

Today we stand at the genesis of the genomics age and are only beginning to realize the exciting opportunities that it brings. To identify remaining obstacles that are in the way of these opportunities, we conducted two surveys. First, we surveyed people with diverse backgrounds and determined factors that deter them from sequencing their genomes. Second, we interviewed researchers at many pharma and biotech companies and identified challenges that they face when working with genomic data.

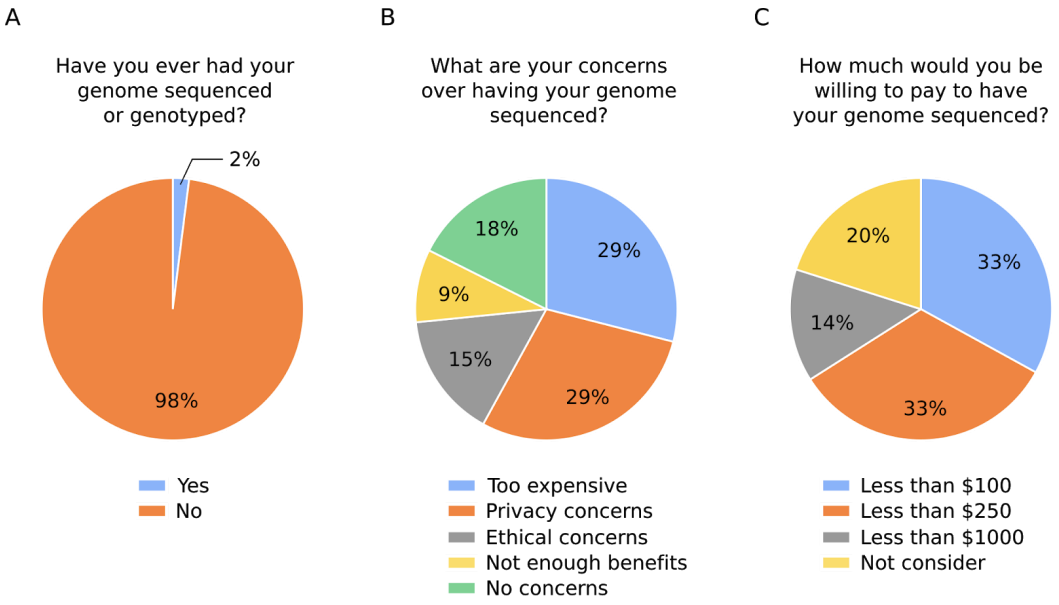

Figure 3. Survey results (sample size = 402).

4.1. Individuals

Only 2% of people who participated in our survey have genotyped or sequenced their genomes (Figure 3A). We identified genome sequencing costs and concerns over genomic data privacy as the main deterrents.

4.1.1. Genome sequencing cost

We found that paying ~ $1000 for personal genome sequencing is considered too expensive by 29% of survey participants (Figure 3B). Of these individuals, a third would sequence their genomes if it cost less than $250, while another third will not pay more than $100 (Figure 3C).

4.1.2. Genomic data protection

29% of survey participants are concerned about genomic data privacy (Figure 3B). We found that people want to keep their data private and share it only when they approve of the

intended use. Yet many personal genomics companies require individuals to completely relinquish ownership of their genomic data.

4.2. Industry

We found that researchers and pharma and biotech companies are constrained by low genomic data quantities, uncertain quality of phenotypic data and inefficient data acquisition logistics. Concerns over resources needed for genomic big data are also increasingly being voiced.

4.2.1. Genomic data

The availability of genomic data is low because very few people have sequenced their genomes to date (Figure 3A). This is a big problem, because very large genomic datasets are needed to establish correlations between genetic variants and traits, such as disease predispositions. This is the case because most traits are the product of complex interactions of many genetic variants, whereby the effects of individual genetic variants are on average very small. To study the effects of millions of genetic variants, machine learning, in particular deep learning, is increasingly being used. Deep learning has been applied with great success to several fields, including computer vision and natural language processing, in part due to an abundance of data to train models. The field of genomics still lacks the data necessary to become another deep learning success story.

4.2.2. Phenotypic data

Phenotypic data refers to information about all personal traits including history of disease. Phenotypic data is used together with genomic data to identify associations between genetic variants and traits. In regards to phenotypic data, buyers are looking for three things: First, data buyers are usually not interested in random datasets, but instead seek to acquire genomic data from individuals with specific phenotypes, such as particular medical conditions. Second, individuals who share their genomic data must also be willing to provide phenotypic data, since without it the genomic data is not particularly useful. Third, quality of collected phenotypic data is often uncertain, because it is typically collected through middlemen, typically personal genomics companies, which rely on self-reported data.

4.2.3. Data acquisition

Our survey has shown that pharma and biotech companies acquire genomic data from many different for-profit (e.g. 23andMe) and non-profit (e.g. UK Biobank) databanks. Hereby buyers encounter two challenges: First, data acquisition logistics lack automation and are therefore very inefficient. Signing contracts, making payments and transferring data requires manual labour and slows down data acquisition. Second, genomic and phenotypic data acquired from different sources is often encoded using different data formats and standardizing across

diverse datasets is very time consuming. As a result, pharma and biotech companies spend a considerable amount of time aggregating and curating across datasets.

4.2.4. Genomic big data

Although relatively few human genomes have been sequenced to date (~ 1 million according to some estimates [35]), the challenges of genomic big data have already begun to emerge, since sequencing of a single human genomes typically produces ~ 200 gigabytes of data that must be processed using compute-intensive algorithms. According to estimates, 100 million to 2 billion human genomes will be sequenced by 2025 [36]. This will create three challenges. First, many exabytes of disk space will be required to store genomic big data. Second, network transfer speeds will limit data sharing. Third, processing and analysis of genomic big data is projected to take trillions of CPU hours [36].

5. The Nebula model

The traditional business model of direct-to-consumer personal genomics companies is illustrated in Figure 4. People pay to sequence or genotype their genomes and receive analysis results. Personal genomics companies keep the genomic data and sell it to pharma and biotech companies that use the data for research and development. This model addresses none of the challenges detailed in the previous sections.

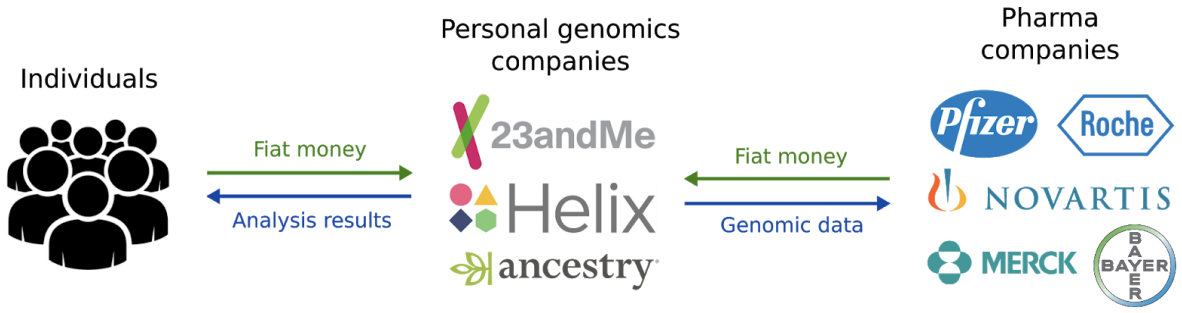

Figure 4. Traditional business model of personal genomics companies.

The Nebula model, shown in FIgure 5, eliminates personal genomics companies as middlemen between data owners and data buyers. Instead, data owners can acquire their personal genomic data from Nebula sequencing facilities or other sources, join the Nebula blockchain-based, peer-to-peer network and directly connect with data buyers. As detailed in the following sections, this model reduces effective sequencing costs and enhances protection of personal genomic data. It also satisfies the needs of data buyers in regards to data availability, data acquisition logistics and resources needed for genomic big data.

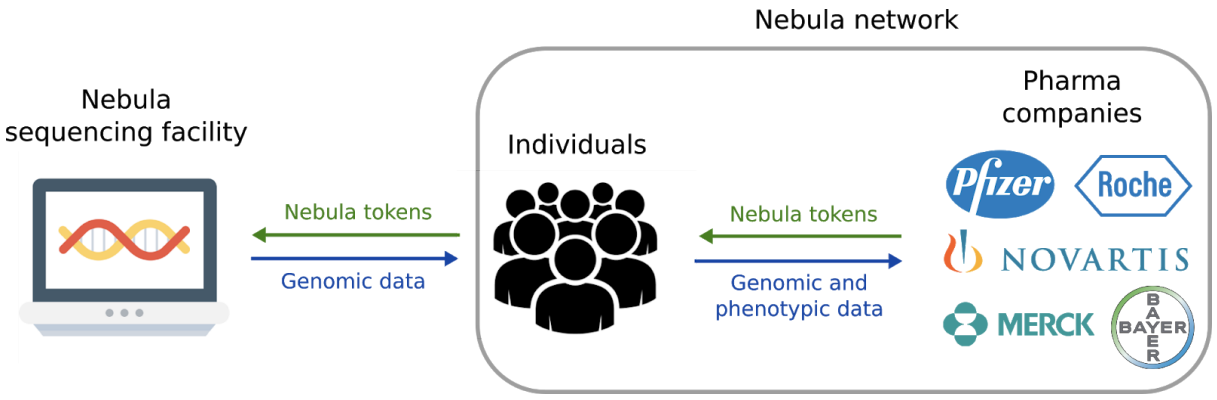

Figure 5. The Nebula model.

5.1.1. Lower sequencing costs

Nebula reduces effective sequencing costs in two ways. First, individuals who have not yet sequenced their personal genomes can join the Nebula network and participate in paid surveys. Thereby data buyers can identify individuals with phenotypes of interest, such as particular medical conditions, and offer to subsidize their genome sequencing costs. As sequencing technology advances and sequencing costs decrease, buyers will be increasingly able to fully pay for personal genome sequencing of many people. Second, individuals who acquired their personal genomic data from Nebula sequencing facilities or other personal genomics companies, can join the Nebula network and profit from selling access to their data. Lowering sequencing costs will incentivize more people to sequence their genomes and result in growth of genomic data that will fuel medical research.

5.1.2. Enhanced data protection

On the Nebula network personal data is protected through multiple mechanisms. First, data owners privately store their personal genomic and phenotypic data and control who can access it. Second, shared data is encrypted and securely analyzed using Intel Software Guard Extensions (SGX) and homomorphic encryption. Data buyers never see personal genomic data in plaintext. Third, while sharing data and receiving payments, data owners remain pseudo-anonymous. Nebula network addresses are cryptographic identifiers that are not associated with any personal information. Fourth, data buyers are required to be fully transparent about their identity, which is verified by Nebula Genomics, and all data transaction records are immutably stored in the Nebula blockchain.

5.1.3. Efficient data acquisition

The Nebula network enables data buyers to efficiently acquire large genomic datasets. First, decentralized private data storage helps address fragmentation of genomic data across

data silos. People who own their personal genomic data and organizations that own genomic databanks can offer their data on the Nebula network while retaining ownership of the data. Thus large-scale data aggregation is achieved by avoiding centralization of data storage. Additionally, sequencing subsidies enable data buyers to actively generate genomic data instead of passively relying on existing datasets. Second, by enabling data buyers to connect directly with data owners, acquisition of high quality phenotypic data will be greatly facilitated. The Nebula smart contract-based survey tool will enable data buyers to design surveys that consist of interdependent questions and elicit true responses. Furthermore, data buyers can directly message individual data owners and ask any specific questions about their phenotype. Third, we will define standard formats for genomic and phenotypic data that is offered on the Nebula network. Nebula Genomics founders' extensive experience with the Harvard Personal Genome Project (personalgenomes.org) and Veritas Genetics has allowed them to gather unique expertise on how to create compatibility among isolated datasets. Data standardization will help data buyers curate large and diverse datasets and save significant costs. Fourth, through utilization of smart contracts, Nebula will facilitate data purchases. This will significantly accelerate data acquisition by automating the process of signing contracts, making payments, and transferring data.

5.1.4. Big data ready

The Nebula network is primed for the forthcoming explosion of genomic data. First, the storage space needed for genomic big data is created by allowing data owners to privately store their data. Thus, Nebula uses available storage space on the edges of network and thereby eliminates the need for centralized data storage. Second, to enable buyers to compute on genomic data, Nebula introduces space efficient data encoding formats that enable rapid transfers of genomic data summaries over the network. Third, available computing resources are leveraged by enabling buyers to utilize any computing hardware that supports Intel Software Guard Extensions (SGX). Hence, data can be analyzed on compute nodes provided by Nebula Genomics, data buyers themselves, or any third party.

6. The Nebula economy

The growth of the Nebula network will be driven by three forces: technological advancement, industry needs, and the consumer market. Figure 6A illustrates these relationships. First, as DNA sequencing technology advances, the costs of human genome sequencing will further decrease and more people will sequence their genomes. Second, industry demand for genomic data will attract more people to join the Nebula network and share their data, thereby further increasing the value of the network and resulting in even more demand. Third, genomic data growth will enable medical research that will expand our knowledge of human genetics,

thereby enabling better interpretation of individual human genomes and ultimately attracting more people to sequence their genomes.

Nebula tokens will be the currency of the Nebula network. The growth of the Nebula network will set in motion a circular flow of Nebula tokens as illustrated in Figure 6B. Individuals will buy personal genome sequencing at Nebula sequencing facilities and pay with Nebula tokens, data buyers will use Nebula tokens to purchase access to genomic and phenotypic data, and Nebula Genomics will sell Nebula tokens to data buyers for fiat money.

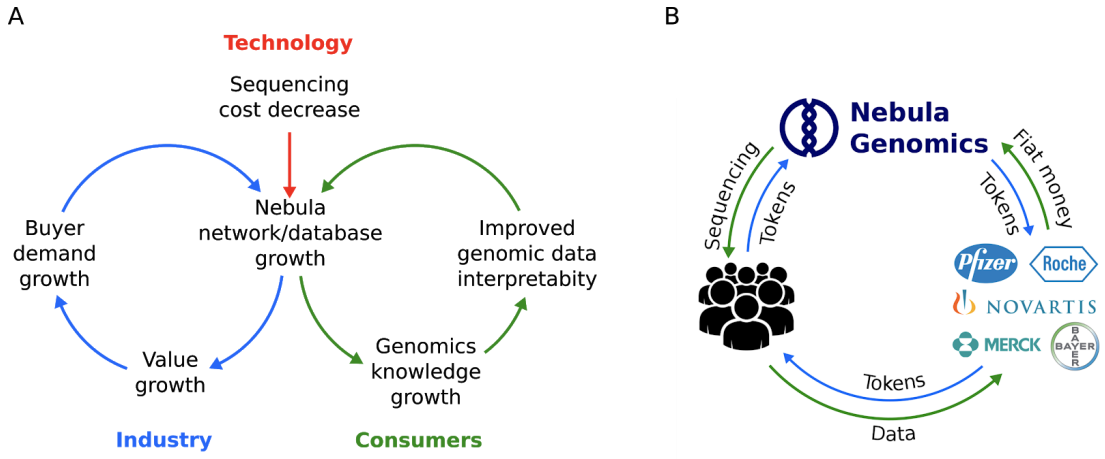

Figure 6. (A) Growth of the Nebula network. (B) Circular flow of Nebula tokens.

7. Personal genomics companies in comparison

	Nebula Genomics	23andMe	ancestry	Helix	Veritas	genos	EncrypGen
Important to consumers: Private storage of personal data	✓	✗	✗	✗	✗	✗	✓
Important to consumers: Data protection through secure computing	✓	✗	✗	✗	✗	✗	✗
Important to consumers: Data buyers subsidize sequencing costs	✓	✗	✗	✗	✗	✗	✗
Important to consumers: Data owners are paid for their data	✓	✗	✗	✗	✗	✓	✓
Whole genome sequencing	✓	✗	✗	✗	✓	✗	✗
Important to pharma/biotech: Data aggregation and standardization	✓	✗	✗	✗	✓	✗	✗
Important to pharma/biotech: Fast acquisition of large datasets	✓	✗	✗	✗	✓	✗	?
Important to pharma/biotech: Data buyers can directly contact data owners	✓	✗	✗	✗	✗	✗	?
Important to pharma/biotech: Big data ready	✓	?	?	?	✓	?	?
Support of third-party apps	✓	✗	✗	✓	✗	✗	✗

Figure 7. Nebula Genomics compared to other personal genomics companies.

8. Technical implementation

8.1. Nebula network

The Nebula network is built on the Blockstack (blockstack.org) platform and the Ethereum-derived Nebula blockchain. Figure 8 gives an overview of the functionality of the Nebula network. The different functionalities labeled with circled letters will be described in detail throughout the technical section.

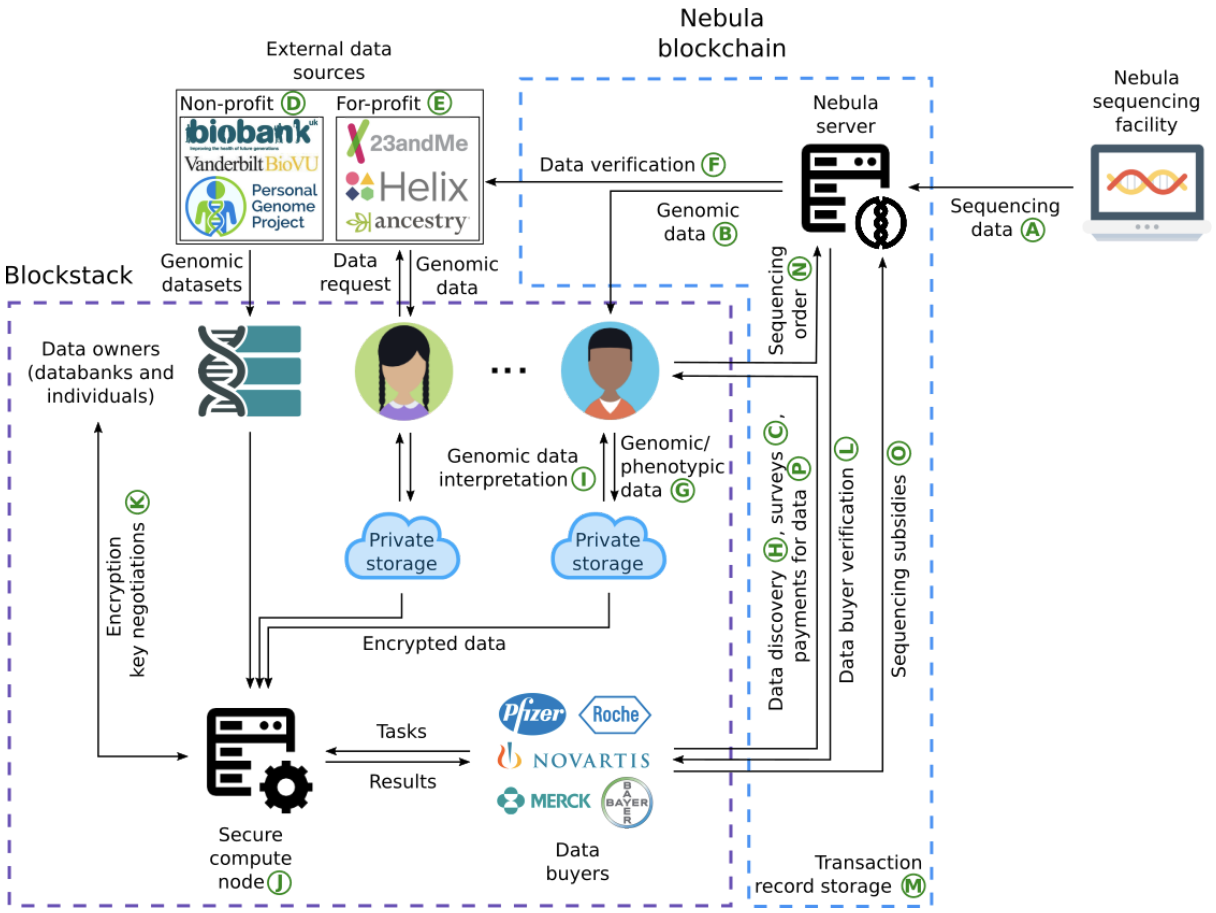

Figure 8. Overview of the Nebula network.

The Nebula network consists of data owner nodes, data buyer nodes, secure compute nodes, and Nebula servers.

1. **Data owner nodes** belong to individuals who want to share their personal genomic and phenotypic data or organizations that own genomic databanks. Data owner nodes utilize Blockstack to privately store their data.

2. **Data buyer nodes** typically belong to pharma and biotech companies. Data buyer nodes purchase genomic and phenotypic data from data owner nodes with Nebula tokens and analyze the data on secure compute nodes. Furthermore, data buyer nodes can subsidize sequencing costs of selected data owners and send out survey questions to generate phenotypic data.
3. **Secure compute nodes** run the Arvados (arvados.org) bioinformatics platform to compute on genomic data. Data privacy is protected through utilization of Intel Software Guard Extensions (SGX) and partially homomorphic encryption. Secure compute nodes can be operated by Nebula Genomics, data buyers, or any third party.
4. **Nebula servers** process sequencing data generated at Nebula sequencing facilities, validate third-party genomic data, and verify identity of data buyers. Importantly, these are mostly functions that are not strictly necessary for the operation of the Nebula network. Thus the Nebula network does not strongly depend on any central authority.

8.2. Genomic data generation

8.2.1. Human genome sequencing

At Nebula sequencing facilities, samples will be sequenced using next-generation DNA sequencing (Figure 8A). Next-generation sequencing of a human genome generates billions of short reads of up to ~ 250 letters in length. Because sequencing is error-prone, each letter of the genome is typically sequenced multiple times. A typical personal genome sequencing file contains ~ 1 billion sequencing reads and is ~ 150 to 200 gigabytes in size.

Nebula Genomics will sequence its samples in in partnership with Veritas Genetics. Veritas Genetics operates CLIA-certified laboratories and has invested $10 million in the latest Illumina NovaSeq high-throughput DNA sequencing systems. By partnering with Veritas, Nebula Genomics will avoid the costs and regulatory hurdles associated with setting up and operating a certified DNA sequencing facility.

8.2.2. Sequencing data processing

Sequencing data generated at Nebula sequencing facilities will be processed on Nebula servers (Figure 8B). First, the genome sequence will be reconstructed by aligning the sequencing reads to a reference human genome and then the present genetic variants will be identified. To enable fast data transfers over the network, the encoding of variant lists must be space-efficient. Furthermore, the encoding scheme must support efficient computations, particularly for machine learning. To this end, we will use an encoding scheme called genome tiling [37] (Figure 9). Hereby the genome is partitioned into overlapping, variable-length sequences and each tile is uniquely represented by a hash digest of the sequence it contains. Tile variants at all tile positions are collected in a tile library that grows as new genomes are sequenced and new variants are

discovered. Individual genomes are represented as arrays of sequence hashes referencing the tile library. These hash arrays are transferred to data owner nodes and can later be shared with data buyers. Since human genomes represented by hash arrays are only ~ 10 megabytes in size, they can be frequently and rapidly transferred over the network. Files that store aligned sequencing reads are transferred to data owner nodes as well. These files are large, 150 to 200 gigabytes, but must be transferred only once from Nebula servers to data owner nodes. They will not be shared with buyers, but data owners can potentially extract additional information from them in the future. After file transfer is completed, all data is deleted from the Nebula servers.

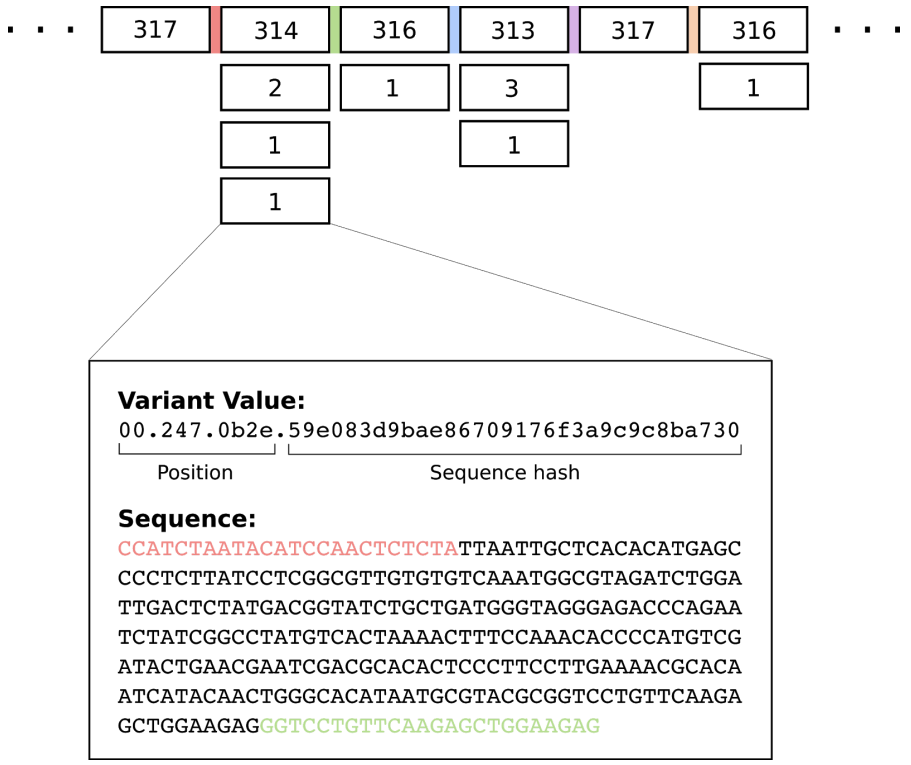

Figure 9. A section of a tile library. The rectangles represent tile variants. The numbers in the rectangles are observed variant frequencies. Colored stripes are unique sequences between tiles.

8.3. Phenotypic data generation

To generate phenotypic data, buyers will be able to send survey questions to data owners (Figure 8C). Importantly, surveys can be designed as chains of interdependent questions that will reveal the true phenotypes even if the survey participants are incentivized to give false responses to make their data appear more valuable. Figure 10 illustrates a survey design that would produce accurate responses by requiring survey participants to provide matching answers about their symptoms, prescribed medication, and diagnosis. Additionally, it will be possible to evaluate the correctness of survey responses on a population level since the prevalence of common medical conditions is well known.

A phenotype data standard for the harmonization of phenotypes across databases is currently being developed by a consortium of biobanks (e.g. UK Biobank and Vanderbilt BioVU) and other organizations housing large-scale genomic databanks. Nebula Genomics and its partners are participating in this early-stage effort to allow efficient use of available data.

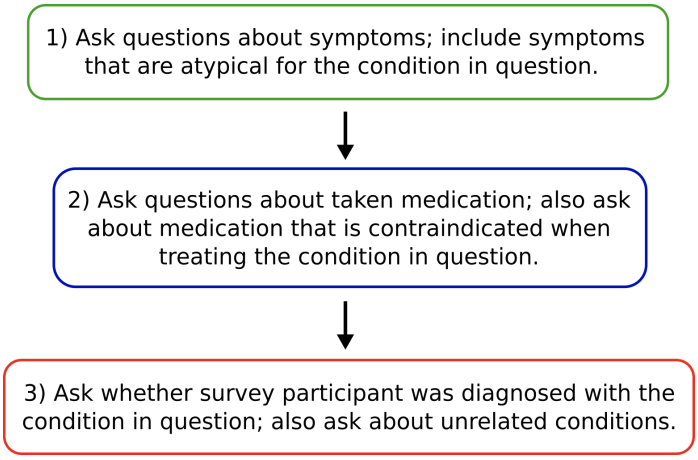

Figure 10. Multi-stage survey design to elicit true responses to survey questions.

8.4. External data aggregation

Genomic data that was not generated at Nebula sequencing facilities can be offered on the Nebula network as well. This includes personal genomic data as well as data from non-profit (8D) and for-profit (8E) genomic databanks. First, data owners will use a tool provided by Nebula to bring their data into the genome tile format. Second, Nebula servers will verify the authenticity of the data (Figure 8F). To this end, owners of of personal genomic data will be required to declare the source of their data and provide proof that the data is real, such as a screenshot of personal genomic company user account page. Similarly, organizations that want to offer data from their genomic datasets on the Nebula network will have to be verified by Nebula Genomics staff. Importantly, verification of third-party data is not strictly necessary and data owners may choose to offer their data without verification. The market will determine whether buyers are willing to pay for unverified data.

8.5. Data storage

8.5.1. Genomic and phenotypic data

Data storage and access control will be implemented using the Blockstack platform for decentralized apps. The Blockstack storage system allows users to choose their own storage providers, such as Dropbox, and regulate access to their data. In the Nebula network data owners will store encrypted hash arrays representing their personal genomes and encrypted text files containing survey responses in their personal cloud storage (Figure 8G). Blockstack also supports

data discovery, which enables implementation of a phenotype registry. Hereby data buyers will be able to query data owner nodes, browse through past surveys, and identify data owners that have answered specific survey questions (Figure 8H).

8.5.2. Genome tile library

The tile library that is referenced by hash arrays representing individual genomes will be stored in public storage such as IPFS/Filecoin (filecoin.io) or BitTorrent. All nodes in the Nebula network will be able to access the tile library. In particular, compute nodes will access the tile library during data analysis.

8.6. Personal genome interpretation

As shown in Figure 8I, data owners will able to interpret their personal genomic data without sharing it with any third party. In this regard, the Nebula variant interpreter will be a classical Blockstack distributed app that is executed locally on user owned data. The first version of the Nebula variant interpreter will be based on the Veritas variant interpreter. However, as the Nebula database grows, we will be able to discover new associations between genetics and health and this information will be used to improve the Nebula variant interpreter. This is important, since a good variant interpreter will incentivize individuals who obtained their genomic data elsewhere to join the Nebula network. This will eventually set in motion a self-reinforcing cycle of data growth and continuously increasing interpretability of genomic data.

8.7. Secure computations

Genomic and phenotypic data will be processed and managed using the Arvados (arvados.org) open source bioinformatics platform, which was designed for genomic and other large-scale biomedical data. Arvados was developed and is supported by Veritas Genetics, and is used by many large institutional customers, including the Wellcome Trust Sanger Institute, IBM Watson, and Harvard Medical School. To enable secure computations, Arvados will be adapted to run inside Intel Secure Guard Extension (SGX) enclaves on secure compute nodes (Figure 8J).

8.7.1. Intel Software Guard Extensions (SGX)

Intel Software Guard Extensions (SGX) is a set of instruction codes that extends Intel's x86 architecture and allows creation of private memory regions, called enclaves. Code and data within enclaves are isolated and protected from external processes. Thus SGX enables secure remote computations on private data by untrusted parties. Intel processors starting with the Skylake microarchitecture support SGX. Importantly, SGX enables secure computations that are several orders of magnitude more efficient than computations on homomorphically encrypted data and secure multi-party computations. It has been demonstrated that SGX can be used to securely compute on genomic data [38,39].

The high level functionality of the SGX framework is illustrated in Figure 11. First, the compute node will create a SGX enclave and launch an Arvados instance within the enclave. A cryptographic signature of the enclave's content will be used to remotely attest enclave integrity. If attestation is successful, data owner nodes will negotiate cryptographic keys with the enclave (also see Figure 8K). Next, data owner nodes will use these keys to encrypt their genomic and phenotypic data and the data will be transferred to the compute node. The compute node will then load the encrypted data into the SGX enclave, where it will be decrypted and analyzed. A second cryptographic key is negotiated between the the SGX enclave and a data buyer node. Finally, the analysis results are encrypted and transferred to the data buyer node where they are decrypted. Note that the compute node outside the SGX enclave does not have access to the input data or the analysis results in plaintext while the data buyer node only gets to see the analysis results.

The ability to execute secure computations by untrusted parties enables flexible compute node designation. Compute nodes can be provided by Nebula Genomics, data buyers themselves, or any third party. The Microsoft Azure cloud computing service recently introduced SGX support, and Google Cloud and AWS will likely follow soon. Furthermore, a distributed secure computing platform based on Intel SGX is currently being developed by Enigma (enigma.co). Nebula Genomics has established a partnership with Enigma.

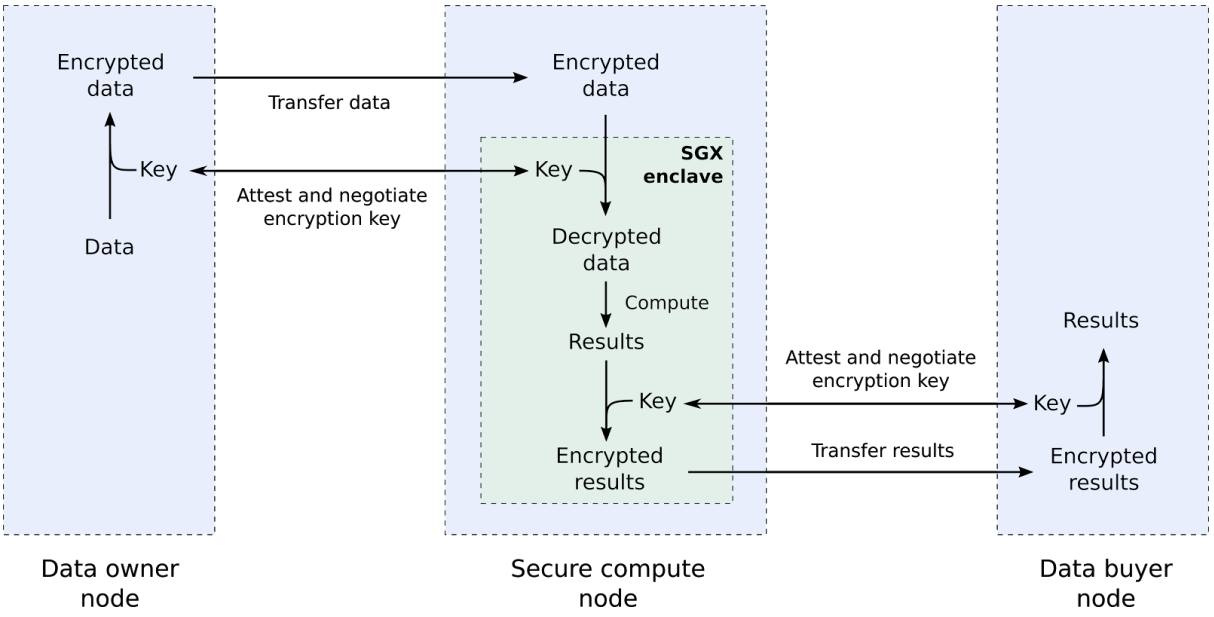

Figure 11. Secure computations with Intel SGX.

8.7.2. Partially homomorphic encryption

A hybrid approach that combines SGX with partially homomorphic encryption can significantly speed up certain computations [39]. In the Nebula network, data owners encrypt and

share personal genomic and phenotypic data with secure compute nodes. Hence the number of encrypted data files that must be decrypted by a SGX enclave corresponds to the number of data owner nodes that shared their data. This can introduce a significant computational overhead.

The first step in many bioinformatics computations is the generation of contingency tables that contain counts of observed genomic variants and corresponding phenotypes. Since computing of contingency tables requires only addition operations, computations can be executed on data encrypted with an additively homomorphic encryption scheme like the Paillier cryptosystem. This is particularly useful for queries that selectively test the significance of selected genomic variants. First, each data owner node encrypts the value 1 or 0, representing presence or absence of a genomic variant, using an additively homomorphic encryption scheme. Then, a compute node can sum all encrypted values outside the SGX enclave. The encrypted sum can then be decrypted inside the SGX enclave and further computations can be performed. Thus, additively homomorphic encryption reduces the number of required decryptions to one. Figure 12 illustrates the hybrid approach of combining SGX with additively homomorphic encryption.

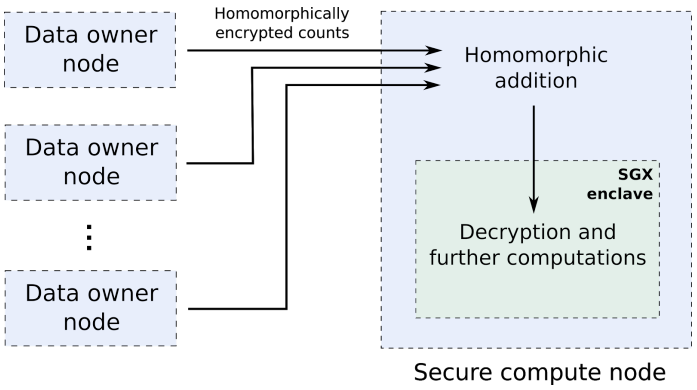

Figure 12. Hybrid approach combining SGX and additively homomorphic encryption.

8.7.3. SGX-GPU hybrid computations

Using Intel SGX has two major drawbacks. First, software must be carefully designed to run inside SGX enclaves without leaking any private data to the outside. We expect that adapting Arvados to run inside SGX enclaves will not be an easy task. Second, all computations must be executed on Intel CPUs, which means that computations can not be GPU accelerated. This is a particularly severe restraint for machine learning algorithms that greatly benefit from GPU acceleration. This problem must be addressed, since machine learning plays an increasingly important role in bioinformatics.

We propose a hybrid approach of data protection in SGX enclaves and GPU accelerated computations (Figure 13). Hereby data will be aggregated and preprocessed in SGX enclaves while the computationally intensive calculations will be executed by GPUs outside the enclaves.

The preprocessing in SGX enclaves will protect data privacy in three ways. First, all data will be perfectly anonymized - SGX preprocessing will hide the source of the input data. Second, only aggregate data summaries, like contingency tables, will output from a SGX enclave. Hash arrays encoding whole genomes will never be exposed. Third, as an additional security measure, random noise can be added to the data. This is a common method that is used to achieve differential privacy of genomic data [40]. An additional advantage of the SGX-GPU hybrid model is that the whole complexity of Arvados can be kept outside of SGX enclaves. This significantly reduces the required engineering effort, since only a relatively simple data preprocessor must be built to run inside SGX enclaves.

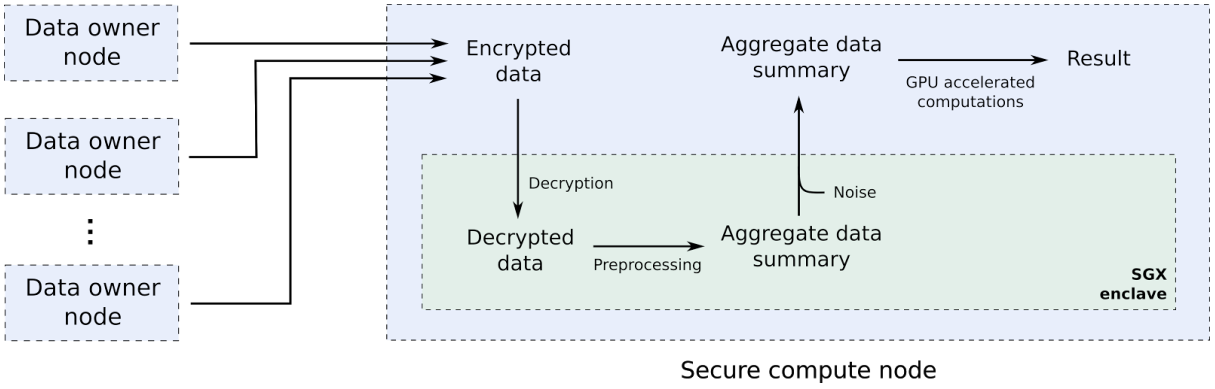

Figure 13. SGX-GPU hybrid computations.

8.8. Data owner protection

While secure computations are crucial to data protection, additional mechanisms are necessary to protect data owners. In particular, under some circumstances data owners may have to reveal their phenotypic information in plaintext. For instance, a data buyer who considers subsidizing the sequencing of a data owner will request to see the unencrypted phenotypic data.

8.8.1. Pseudo-anonymity

The Ethereum blockchain offers pseudo-anonymity to data owner nodes. Network addresses are cryptographic identifiers that are not associated with any personal information.

8.8.2. Buyer verification

Most data owners would want to know who they are sharing their data with. Thus data buyers should be fully transparent about their identity. To be verified by Nebula servers, as illustrated in Figure 8L, buyers will be required to provide their real names and company affiliation as well as legally agree not to share acquired data with any undisclosed third party. Nebula staff will verify this information and Nebula servers will whitelist network addresses of verified buyers. Hence data owners will be able to verify buyer identity by contacting Nebula servers.

8.8.3. Transaction records

As indicated in Figure 8M, all data-sharing records will be immutably stored in the Nebula blockchain. Thus, if a dataset is found to have been released without permission, it might be possible to identify the buyer of that data.

8.9. Payment

8.9.1. Sequencing

Nebula sequencing facilities will offer whole genome sequencing for ~ $1000 and exome sequencing for ~ $300 (Figure 8N). The prices will be in Nebula tokens and will decrease at an exponential rate as DNA sequencing becomes cheaper. As mentioned previously, data buyers will be able to subsidize sequencing cost of selected individuals. This functionality will be implemented with a smart contract. Hereby a buyer node will transfer Nebula tokens to the Nebula server and specify the Nebula network address of a data owner who will be offered subsidized sequencing (Figure 8O). Importantly, as sequencing prices continue decreasing, buyers will be increasingly able to fully subsidize sequencing costs.

8.9.2. Surveys

The Nebula survey design tool will utilize Ethereum smart contracts and enable data buyers to create highly customized surveys. For example, data buyers may choose to pay all survey participants an equal amount of Nebula tokens or alternatively define different token amounts that will be awarded for different combinations of responses. For example, if a survey participant is found to be affected by a condition that is of interest to the data buyer, the highest token reward will be automatically paid out. Responses that suggest that the survey participant is not affected by the condition in question will trigger a lower token payment. Contradictory responses indicating dishonesty will not be rewarded.

8.9.3. Data

Buyers will purchase access to personal genomic data using Ethereum smart contracts (Figure 8P). Hereby, data owners will receive token payments and, in turn, their encrypted genomic data will be transferred to compute nodes for secure computations. Purchases of phenotypic data will be handled similarly. Importantly, buying existing phenotypic data will be cheaper for buyers, since it will not require data owners to participate in a new survey.

At Nebula Genomics our goal is to promote scientific advancement to the benefit of all. In particular we want to support research conducted by non-profit institutions, such as universities. To this end we will subsidize data purchases of buyers with verified academic affiliations. We will implement this through a smart contract that will trigger a token transfer from a Nebula server to data owner nodes whenever a buyer with verified academic affiliations purchases data.

8.10. Third-party apps

The concept of an app store for applications that interpret genomic data was recently pioneered by Helix (helix.com). Helix sequences customer samples and third party developers build apps that provide interpretation. This is an interesting model, but it suffers from the drawbacks of centralization. First, data and apps are siloed on Helix servers. Second, the helix ecosystem is not open to data that was generated elsewhere and customers are furthermore limited to apps that were approved by Helix. We believe that Nebula will be a superior platform for third party apps because it creates a decentralized, open ecosystem that aggregates genomic data from all available sources. Developers can build decentralized Blockstack apps that will run locally on data owner computers and interpret privately stored genomic data. Thus third party apps will not violate genomic data privacy and the open ecosystem will eliminate entry barriers for app developers.

9. Future directions

9.1. Decentralized sequencing

By enabling decentralized data storage and secure computations, Nebula will set a new standard for genomic data protection. However, as shown in Figure 7A, data generation still occurs at centralized sequencing facilities. If the security of a sequencing facility is compromised, biological samples or genomic data may be stolen. This privacy risk can only be addressed if sequencing itself is decentralized. Ideally, instead of entrusting biological samples to a sequencing provider, individuals should be able to purchase a DNA sequencing machine and sequence their samples themselves. Today, commonly used DNA sequencers are large, difficult to operate, and cost up to $1 million. However, innovation is occurring rapidly. Novel, groundbreaking DNA sequencing technologies make it possible to build DNA sequencers that fit in the palm of one's hand, do not require complicated sample preparation and cost less than $1000. However, these technologies are still maturing and are therefore not yet ready for the consumer market. Nebula Genomics will stay at the cutting edge of sequencing technology and seek to enable affordable, private personal genomics.

9.2. Personal Health Records

The described implementation of the Nebula platform supports storage and analysis of genomic data and structured phenotypic information. Our long term goal is to support more diverse, unstructured phenotypic data. In particular, we wish to support medical information such as drug prescriptions and medical images as well as data collected from activity trackers and other devices. We envision Nebula to become a Personal Health Records (PHR) system that derives valuable insight into personal health by integrating genomic data with various phenotypic information.

10. References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
2. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
3. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: https://www.genome.gov/sequencingcostsdata/. (Accessed: 11th January 2018)
4. Illumina Introduces the NovaSeq Series — a New Architecture Designed to Usher in the $100 Genome. (2017). Available at: https://www.illumina.com/company/news-center/press-releases/press-release-details.html%3Fnewsid%3D2236383.
5. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
6. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
7. Lindor, N. M., Thibodeau, S. N. & Burke, W. Whole-Genome Sequencing in Healthy People. *Mayo Clin. Proc.* **92,** 159–172 (2017).
8. Schwarz, U. I. *et al.* Genetic determinants of response to warfarin during initial anticoagulation. *N. Engl. J. Med.* **358,** 999–1008 (2008).
9. Petrucelli, N., Daly, M. B. & Pal, T. BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1998).
10. Susan Caminiti, S. to C. C. More than 80 million Americans have this deadly disease, and many don't even know it. *CNBC* (2016). Available at: https://www.cnbc.com/2017/10/31/fatty-liver-disease-affects-80-million-americans.html. (Accessed: 25th January 2018)
11. Harper, A. R., Nayee, S. & Topol, E. J. Protective alleles and modifier variants in human health

and disease. *Nat. Rev. Genet.* **16,** 689–701 (2015).

12. Berg, J. S. *et al.* An informatics approach to analyzing the incidentalome. *Genet. Med.* **15,** 36–44 (2013).
13. Levine, D. S. RARE Diseases: Facts and Statistics. *Global Genes* (2012). Available at: http://globalgenes.org/rare-diseases-facts-statistics/. (Accessed: 25th January 2018)
14. Coletta, A. Genetic Profiling for Weight Loss: Potential Candidate Genes. *Bioenergetics* **04,** (2015).
15. Elsamanoudy, A. Z., Neamat-Allah, M. A. M., Mohammad, F. A. H., Hassanien, M. & Nada, H. A. The role of nutrition related genes and nutrigenetics in understanding the pathogenesis of cancer. *Journal of Microscopy and Ultrastructure* **4,** 115–122 (2016).
16. Yang, A., Palmer, A. A. & de Wit, H. Genetics of caffeine consumption and responses to caffeine. *Psychopharmacology* **211,** 245–257 (2010).
17. Mattar, R., de Campos Mazo, D. F. & Carrilho, F. J. Lactose intolerance: diagnosis, genetic, and clinical factors. *Clin. Exp. Gastroenterol.* **5,** 113–121 (2012).
18. Freeman, H. J. Risk factors in familial forms of celiac disease. *World J. Gastroenterol.* **16,** 1828–1831 (2010).
19. Borel, P. & Desmarchelier, C. Genetic Variations Associated with Vitamin A Status and Vitamin A Bioavailability. *Nutrients* **9,** (2017).
20. Delanghe, J. R. *et al.* Vitamin C deficiency: more than just a nutritional disorder. *Genes Nutr.* **6,** 341–346 (2011).
21. Malloy, P. J. & Feldman, D. Genetic disorders and defects in vitamin d action. *Endocrinol. Metab. Clin. North Am.* **39,** 333–46, table of contents (2010).
22. Guth, L. M. & Roth, S. M. Genetic influence on athletic performance. *Curr. Opin. Pediatr.* **25,** 653–658 (2013).
23. O'Connell, K. *et al.* Interactions between collagen gene variants and risk of anterior cruciate ligament rupture. *EJSS* **15,** 341–350 (2015).
24. Tiziano, F. D., 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* **3,** 28 (2016).

25. Bennett, E. R., Reuter-Rice, K. & Laskowitz, D. T. Genetic Influences in Traumatic Brain Injury. in *Translational Research in Traumatic Brain Injury* (eds. Laskowitz, D. & Grant, G.) (CRC Press/Taylor and Francis Group, 2015).
26. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
27. Mendias, C. L. *et al.* Haploinsufficiency of myostatin protects against aging-related declines in muscle function and enhances the longevity of mice. *Aging Cell* **14,** 704–706 (2015).
28. Camporez, J.-P. G. *et al.* Anti-myostatin antibody increases muscle mass and strength and improves insulin sensitivity in old mice. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 2212–2217 (2016).
29. Hirschler, B. Amgen buys Icelandic gene hunter Decode for $415 million. *Reuters* (2012). Available at: http://www.reuters.com/article/us-amgen-decode/amgen-buys-icelandic-gene-hunter-decode-for-415-million-idUSBRE8B90IU20121210. (Accessed: 14th October 2017)
30. Palmer, K. M. *et al.* Why Iceland Is the World's Greatest Genetic Laboratory. *Wired* (2015).
31. Herper, M. Surprise! With $60 Million Genentech Deal, 23andMe Has A Business Plan. *Forbes* (2015). Available at: https://www.forbes.com/sites/matthewherper/2015/01/06/surprise-with-60-million-genentech-deal-23andme-has-a-business-plan/. (Accessed: 1st October 2017)
32. Ledford, H. AstraZeneca launches project to sequence 2 million genomes. *Nature* **532,** 427 (2016).
33. Berkrot, B. Regeneron forms consortium to accelerate gene sequencing project. *Reuters* (2018).
34. Genetics, V. Veritas Genetics Launches Collaboration with Mayo Clinic to Integrate Whole Genome Sequencing into Clinical Care. *PR Newswire* (2018). Available at: https://www.prnewswire.com/news-releases/veritas-genetics-launches-collaboration-with-mayo-clinic-to-integrate-whole-genome-sequencing-into-clinical-care-300578640.html. (Accessed: 24th January 2018)
35. Regalado, A. EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced This Year. *MIT Technology Review* (2014).
36. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13,** e1002195 (2015).

37. Guthrie, S. *et al. Tiling the genome into consistently named subsequences enables precision medicine and machine learning with millions of complex individual data-sets*. (PeerJ PrePrints, 2015). doi:10.7287/peerj.preprints.1426v1

38. Chen, F. *et al.* PRESAGE: PRivacy-preserving gEnetic testing via SoftwAre Guard Extension. *BMC Med. Genomics* **10,** 48 (2017).

39. Sadat, M. N. *et al.* SAFETY: Secure gwAs in Federated Environment Through a hYbrid solution with Intel SGX and Homomorphic Encryption. *arXiv [cs.CR]* (2017).

40. Simmons, S., Sahinalp, C. & Berger, B. Enabling Privacy-Preserving GWASs in Heterogeneous Human Populations. *Cell Syst* **3,** 54–61 (2016).