Privacy-Preserving Health and Genomics Data Marketplace Powered by AI and Blockchain

Lavita Technologies
White Paper, v1.2

March, 2023

1 Overview

1.1 Growth of Cloud Computing and Privacy Issues

Advancements in cloud computing are transforming how data is hosted, shared, and processed. The rapid adoption of cloud technologies has created an opportunity to connect different data silos. As shown in Figure 1, the evolution of computing started with the adoption of centralized mainframes by big businesses for critical applications (e.g., transaction processing, census, etc.) in the 1950s. In the 1980s, with technological advances, PCs became more portable, allowing computing mobility and improved productivity. Moving to the 2000s, a significant shift has been seen from personal computing to mobile and cloud computing, where the latter enabled businesses to cost-effectively utilize elastic computing resources on a pay-as-you-go basis.

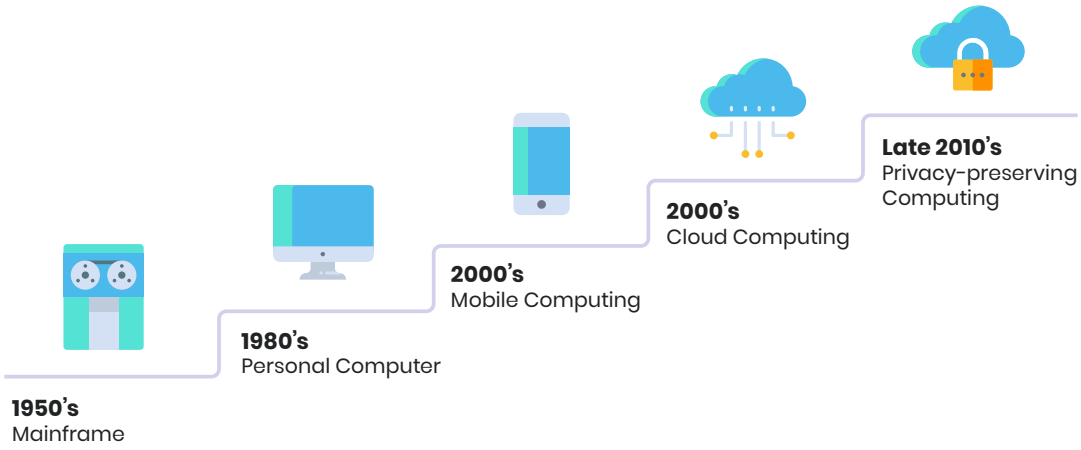


Figure 1: The progression from mainframe to privacy-preserving computing

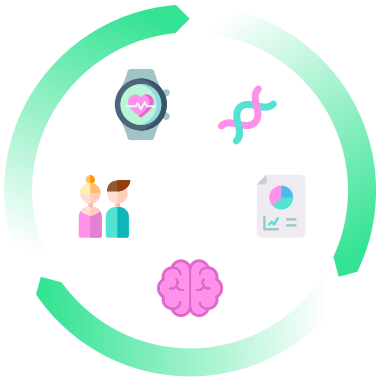
By leveraging the flexibility and scalability of cloud computing, enterprises, small business, as well as, startups can easily increase their productivity and streamline the deployment of innovative products.

Based on Gartner's report, worldwide end-user spending on cloud computing is expected to grow 20.7% in 2023 to a total of $591.8 billion, up from $490.3 billion in 2022, which is nearly $600 billion [1]. There is no doubt that industries (e.g., in the areas of healthcare, finances, internet of things, etc.) are moving towards the cloud-based computing paradigm.

However, security and privacy are becoming one of the major challenges for industries in utilizing scalable and cost-effective cloud-computing technologies. For example, until March 2015, the National Institutes of Health (NIH) did not allow controlled-access genomic data to be uploaded to the public cloud due to privacy concerns. The policy has changed, in response to "the advances made in security protocols" and "the expansion in the volume and complexity of the genomic data". However, the new policy lays the liability for data disclosure entirely on the cloud user side. It was clearly stated in the "NIH Security Best Practices for Controlled-Access Data Submit to the NIH Genomic Data Sharing (GDS) Policy" that cloud users, not the cloud service provider, are responsible for ensuring the security of human genomic data [2]. In the absence of provable and easy-to-use protection methods, this requirement actually deters the use of the cloud. Security and privacy technologies have advanced significantly in the past few years, which increases the feasibility

of privacy-preserving computing. Such a computing paradigm will enable data availability and usability, as well as change the way data is shared in all industries.

Extreme caution must be exercised over data contributorship, privacy and protection; different data silos need to be connected into a seamless and interoperable fabric; proper incentive structure must be introduced for data contributors and data miners built on a blockchain-based infrastructure; as well as privacy-preserving and artificial intelligence(AI)-based data analytic computation must enable collective statistical learning without compromising individual privacy. Figure 2 shows three recent trends driving the need for the next-generation of health applications.



I. Individual Ownership of Data

- **Individual / Decentralized Ownership of Data** is the cornerstone for a fair and functioning decentralized marketplace.
- Smart-contract transactions encompass individual ownership of data while incentivize data owners to share and monetize their data

II. Data Privacy

- **Privacy-preserving Computation** based on developments in Trusted Execution Environment (TEE) and advanced cryptographic technologies allow collective AI and statistical learning over different data silos without compromising individual privacy.

III. Advancements in AI/NLP Technologies

- **Natural Language Processing** and other AI technologies have exponentially expanded into health-related applications in the past decade

Figure 2: Healthcare: confluence of three major trends

1.2 Challenge of Managing Personal Genomic and Health Data

The rapid adoption of electronic health records (EHR) has enabled the meaningful use of healthcare data for advancing biomedical research. Seminal advancement in genomic data generation over the past decade has also impacted biomedical science and related scientific studies. The genesis in data accumulation has made scientific studies on multiple types of medical genomics more realistic [3]. Large and varied biomedical datasets now help researchers understand the relation between our genome codes and our health [4]. In addition, many direct-to-consumer (DTC) genetic testing companies (e.g., 23andMe, Ancestry.com, etc.) have also contributed to the expansion of the market for personal genetic data [5]. According to industry estimates, more than 26 million individuals have had their DNA analyzed by one of DTC genetic testing providers. For example, Ancestry.com and 23andMe have tested more than 14 million and 9 million people respectively, by the end of 2019 [6]. Nowadays, around 1 in 25 American adults have access to personal genetic data. These data will become important resources of data-driven approaches in biomedical science, in particular for precision medicine [7].

Given a large amount of personal genomic data, efficient sharing, proper storage, and rapid processing become critical to achieving these goals. However, various challenges stand in the way of managing, sharing, and processing large-scale genomic data and the associated personal health information. One challenge is the lack of a health data market ecosystem that can democratize data sharing and analysis among data contributors, data miners, and healthcare providers in a

trustworthy manner. As a result, health and genomic data are typically handled in a centralized framework (e.g., dominated by big entities like EHR providers). Individuals own their data but have limited options to choose how their data can be shared and analyzed while appropriately protected and compensated after giving consent. On the other hand, data miners (pharmaceutical companies, research institutions, etc.) usually need to access genomic and healthcare from a large cohort to generate hypotheses and models for developing novel drugs and treatments for diseases. We believe there is a tremendous opportunity and need to develop the first decentralized ecosystem for sharing and analyzing genomic and health data securely and with privacy-preserving primitives.

2 Lavita Platform - A Decentralized Health Data Marketplace

2.1 Leveraging blockchain, AI and privacy-preserving technologies

Lavita platform brings together three technologies, namely distributed blockchain infrastructure, AI, and privacy-preserving computing to implement a fully decentralized data marketplace.

1. Blockchain and smart contracts enable a marketplace that incentivizes data contributors to share their data while protecting their ownership of data. The use of blockchain technology removes the middlemen in the conventional data marketplace, lowers the cost of data curation for medical research, and introduces an incentive mechanism that encourages users' participation to contribute their private data and receive rewards.

   Lavita platform incentivizes individuals and healthcare institutions to share healthcare data, and get rewarded with a new TNT-20 "LAVITA" token, built on the Theta blockchain. This token will be utilized not only for users sharing data but also for offering storage and computation capabilities. This will be enabled by the Theta Edge Network, the decentralized blockchain infrastructure providing secure distributed data storage (Theta EdgeStore), computation, and more. More details of these components are described in Section 3.

2. AI-powered tools ensure that data miners, institutions, and data contributors can find quality data on Lavita and build novel applications on this platform. These tools are implemented by unique techniques for data validation and include a Semantic Search Engine with in-domain knowledge about healthcare and biomedicine. The Semantic Search Engine aims to not only find keywords but determine the intent and contextual meaning of search queries by users and helps them achieve the goal of finding the right information at the right time.

   AI-aided disease prediction has undergone significant development in recent years by improving the diagnosis approaches, supporting clinical decisions, and reducing healthcare costs. Recently, machine learning and deep learning approaches have been applied to clinical predictive modeling with numerous successes [8]. The primary dataset for predictive modeling in the clinical area today is in form of Electronic Health Records (EHR), which offer rich and well-structured information that reflects the disease progression of each patient and is one of the most valuable resources for healthcare analysis. Lavita platform will be able to utilize EHR primary datasets and various other healthcare-related data types (e.g., genomic, survey, clinical) and data modalities (e.g., text, image, audio) combined with AI-powered models and algorithms, including natural language processing (NLP) and natural language understanding (NLU) methods, to build applications to address different needs for individuals and healthcare

institutions. These applications include but are not limited to: accelerating the development of novel treatments for diseases, detecting diseases' symptoms in patients early on, informing patients/individuals with insightful genetic information, research, and the latest findings based on their clinical needs in the platform, or helping clinical trials succeed by connecting eligible individuals with these trials. As part of building AI-powered applications, Lavita leverages the shared de-identified data for training and fine-tuning large language models (LLMs) or various other foundation models [9] without compromising user privacy, confidentiality, or data anonymity. LLMs and foundation models, enabled by *transfer learning* and scaled and powered by better computation, larger and better quality data for training, and novel architectures such as the Transformer [10], have demonstrated promising performance in many tasks including image classification, question answering, document summarization, translating human languages, automatic code generation, and beyond. These models can unlock massive potential in advancing research in biomedicine and address many downstream tasks when pretrained on domain-related corpora (e.g., medicine or science) and fine-tuned on smaller high-quality in-domain datasets (e.g., clinical). Lavita enables the development of powerful AI applications in the biomedical domain by facilitating the creation of LLMs and foundation models and providing high-quality data at scale validated by AI-powered tools.

3. Privacy-preserving computing ensures end user data remains private during the computation phase of statistical machine learning. The key to privacy-preserving computing lies in its power to enable data sharing and analysis among different sub-systems and entities without exposing data to any unauthorized third-party or system.

   The Lavita platform ensures full privacy-preserving computing by implementing a Trusted Execution Environment (TEE) such as Intel Software Guard Extensions (SGX), which is a hardware-based memory encryption system that isolates specific application code and data in memory. As an example, Intel SGX allows user-level code to allocate private regions of memory, known as enclaves, which are designed to be protected from processes running at higher privilege levels. As such, Intel SGX offers a granular level of control and protection to implement privacy. Intel SGX has been proven effective in the context of genomics and health data, as reported by numerous peer-reviewed research publications, including some publications co-authored by Lavita team members and advisors [11, 12, 13, 14].

2.2 Reward for Healthcare Data Contributors and User Workflow

Centralized third-party genetic data companies are getting a high return on the data they own. A preeminent example is 23andMe, which owns the genetic data as well as over 1.5 billion answers to survey questions from over 5 million customers, among which more than 80% consent to research and recontact. These survey questions cover broad aspects of customers' health conditions, including their heart health, inflammation, metabolic, optimal, nutrition, stress management, and family history of health, which can help pharmaceutical companies such as GSK to shift to a "genetics-driven" and "genetics-supported" R&D portfolio, as it has been shown that the drugs designed to target a specific biological mechanism with a strong genetic/biochemical rationale have a twofold higher probability of success in comparison with those without such rationale.

Lavita's decentralized revenue-sharing model is able to retrieve every single data source and directly pay the profit of data usage back to each individual who contributes his/her data. In addition to the free module on-chain services, LAVITA tokens can also be used to purchase the products like

medical insurance once such a plan is plugged in, described more in Section 3.

The user workflow of the Lavita platform is shown in Figure 3, beginning with contributing healthcare data through a fully secure, encrypted mobile app, to aggregating and analyzing data privately and earning rewards in the form of Lavita tokens.

Figure 3: User workflow in the Lavita platform

By utilizing privacy-preserving computation, sensitive data can be protected and controlled during the entire data analysis pipeline which was previously impossible. Such a computing paradigm will prompt data availability and usability, as well as, change the way data is shared among stakeholders. The Lavita platform implements a decentralized, peer-to-peer data marketplace connecting individuals who share private data with the data miners who perform biomedical research.

As shown in Figure 4, the overall process begins with data contributors contributing their own private health data and getting rewarded in LAVITA tokens from data miners. Depending on the type and quality of information that data miners (e.g., pharmaceutical and research labs) utilize, the incentive mechanism will value data differently. See Section 3.3 for more details. These rewards can also be used by individuals to pay for health services to healthcare institutions (e.g., annual health checkups).

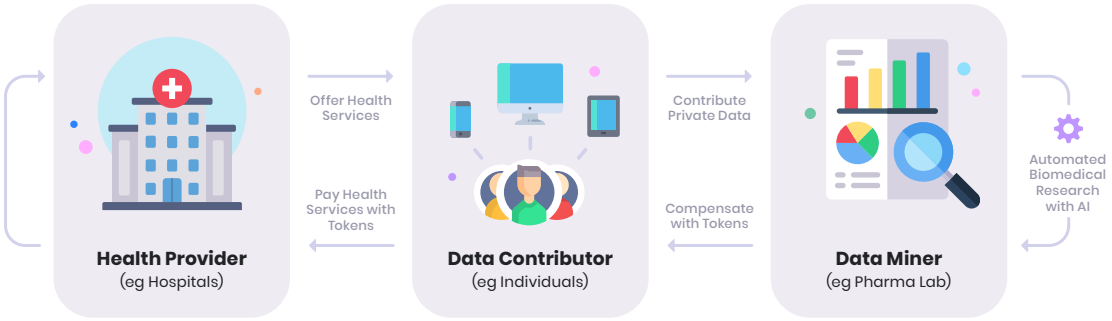


Figure 4: Lavita platform - a decentralized data marketplace

2.3 Five Key Components

The Lavita platform (see Figure 5) consists of five integral components: 1) Blockchain, 2) Secure Data Storage, 3) Secure Computation (e.g., secure hardware, multi-party computation (MPC) [15, 16, 17, 18], homomorphic encryption (HE) [19, 20, 21, 22], zero-knowledge proof (ZKP) [23], and differential privacy), 4) AI-supported Research & Validation Applications (e.g., research engine, data validation tool), and 5) Secure Key Management Services (KMS).

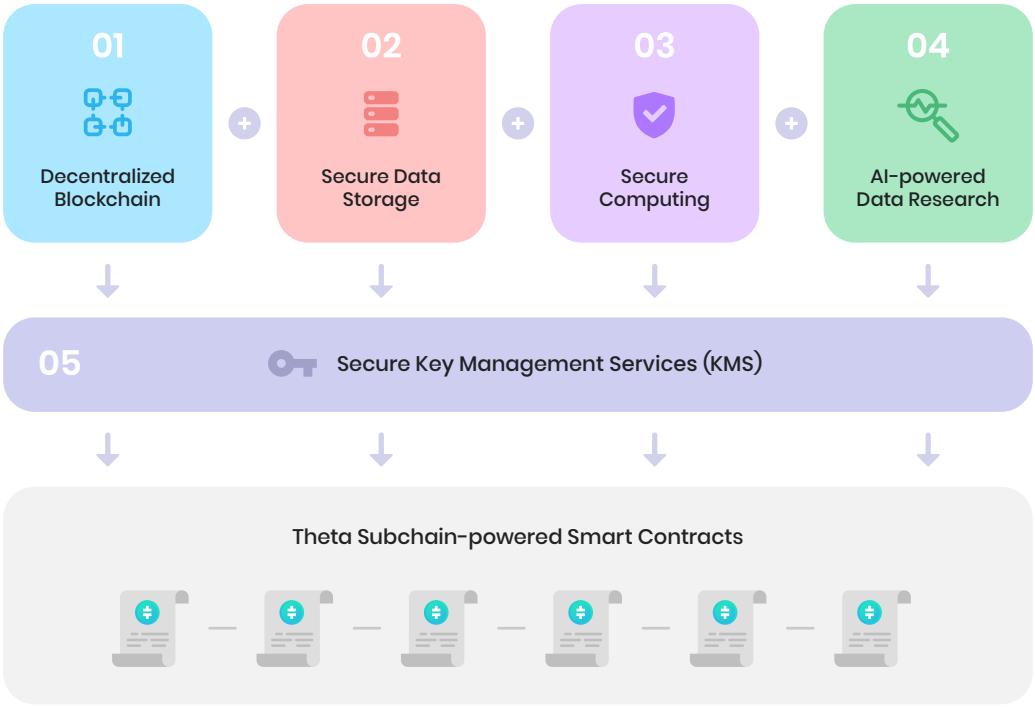


Figure 5: Technological components in the Lavita platform

These five modules enable a trusted, fair marketplace for data contributors and data miners to exchange health data as digital assets. Smart contracts will record transactions on the blockchain upon completing computation and ensure that results are immutable, verifiable, and privacy-preserving. Secure computation is conducted off-chain, while the verification is achieved through distributing trust across multiple parties, instead of relying on a centralized service provider that every participant has to trust.

2.3.1 Decentralized blockchain Powered by Theta Subchain

As shown in Figure 6, all Lavita systems are powered by smart contracts running on a customized Theta Subchain. This custom health data subchain enables full horizontal scalability, without sacrificing privacy and security. Additionally, Lavita plans to fully leverage Theta's decentralized edge network to support two main capabilities: 1) store private healthcare information through Theta EdgeStore, and 2) support secure Lavita workloads. The new TNT-20 token, LAVITA token, will be created on the new subchain to increase the rate of adoption and liquidity of the marketplace through healthcare data contribution, storage, and computation.

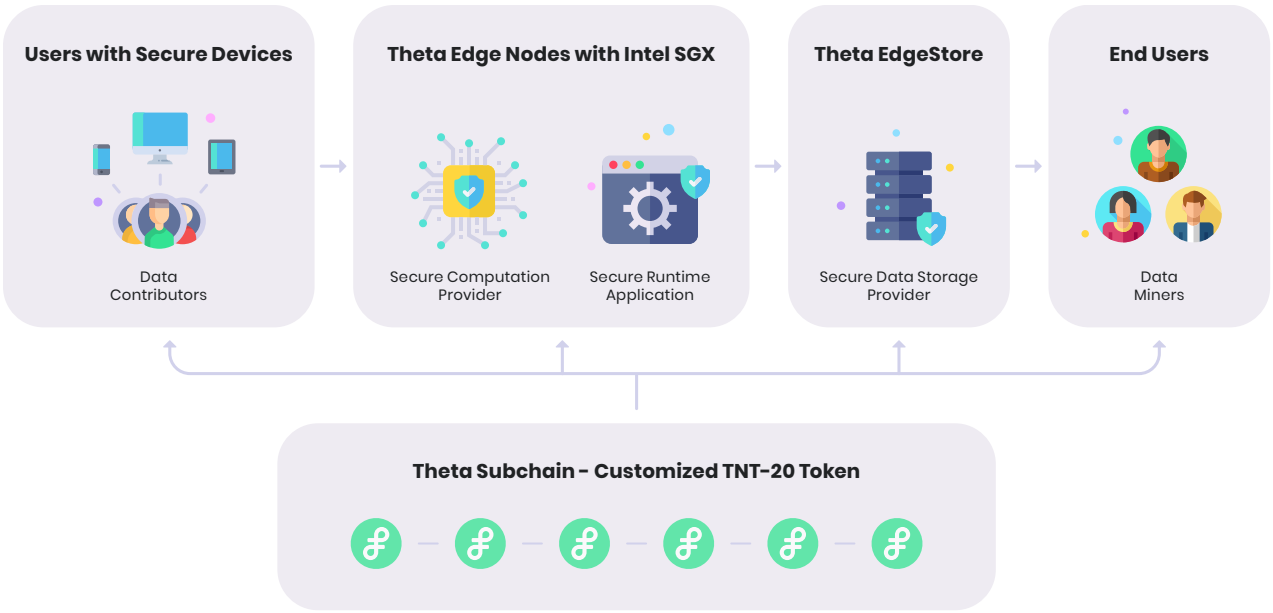


Figure 6: Tokenized data marketplace in Lavita ecosystem

Data Miners are research institutions, clinical institutions, or pharmaceutical companies, whose biomedical research and development require personal genomic data. Data consumers have the incentive to compensate data contributors for sharing their high-quality data. Within the Lavita framework, such compensation is done through the use of smart contracts and blockchain technology.

Data contributors are individuals or organizations such as a biobank or a direct-to-consumer company who are willing to share their genomic data. The data contributor will receive compensation for the data in the form of tokens distributed from the smart contracts.

Secure data storage providers are entities or individuals, who provide decentralized data storage infrastructure via the Theta EdgeStore, for safeguarding the encrypted biomedical data and for offering reliable and scalable access to data upon request.

Secure computation providers are Theta Edge Node operators that run the server-based TEE hardware to provide secure and high-performance biomedical data analysis and computation services.

Secure runtime app providers develop secure runtime applications for data miners to be executed by a secure computation provider, where the secure runtime app can also be hosted on the secure data storage. These apps include intelligent data analytics and direct-to-consumer applications such as deep learning, regression models, association test pipeline, disease risk analysis applications, and so on.

2.3.2 Secure Data Storage Powered by Theta EdgeStore

In the Lavita platform, the data storage service will be supported by a decentralized data storage infrastructure powered by Theta EdgeStore. It provides a high-throughput content-addressed block

storage model in order to achieve better robustness to efficiently handle large-scale genome data from different sources. As shown in Figure 7, data encryption technologies will be adopted for safeguarding biomedical data.

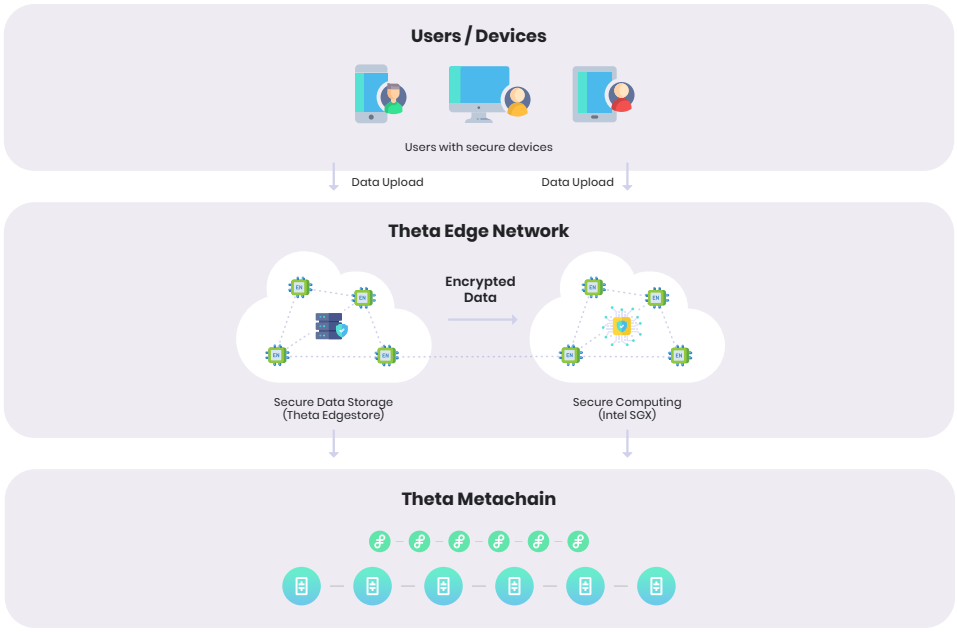


Figure 7: Overview of the secure storage framework in Lavita platform

2.3.3 Secure Computation Leveraging TEE and Theta Edge Network

Trusted execution environment (TEE) provides an isolated memory and computation space within hardware (e.g., Intel sgx enclave), in which sensitive data can be analyzed efficiently and securely. There exists many TEE products in the market, e.g. ARM Trustzone, AMD Secure Encrypted Virtualization and Intel SGX. Upon launch, Lavita will implement Intel server-based SGX (TEE) [13, 24, 12] as our computational node to maximize the security of application code and data, protecting against runtime disclosure or modification. The Secure Computation Workflow is shown in Figure 8.

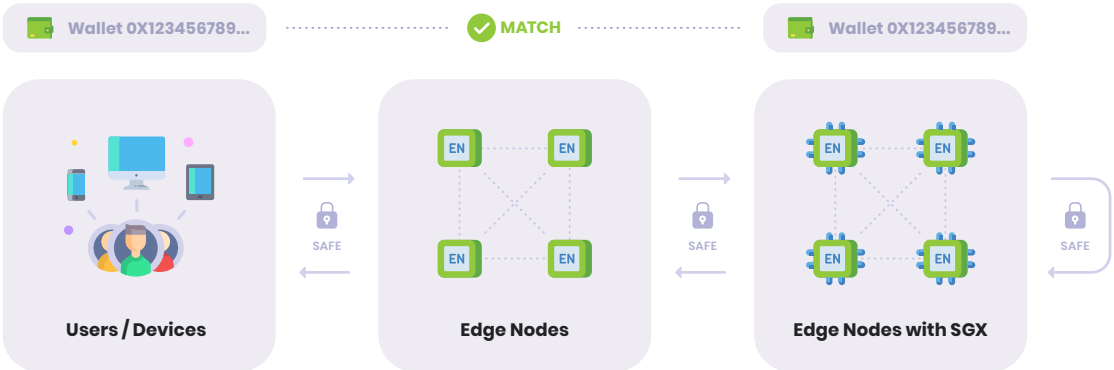


Figure 8: Lavita secure computation workflow

More specifically, the Intel Software Guard Extensions (SGX), a hardware-enabled TEE, can be installed for the operation to achieve secure and efficient computation as shown in Figure 9.

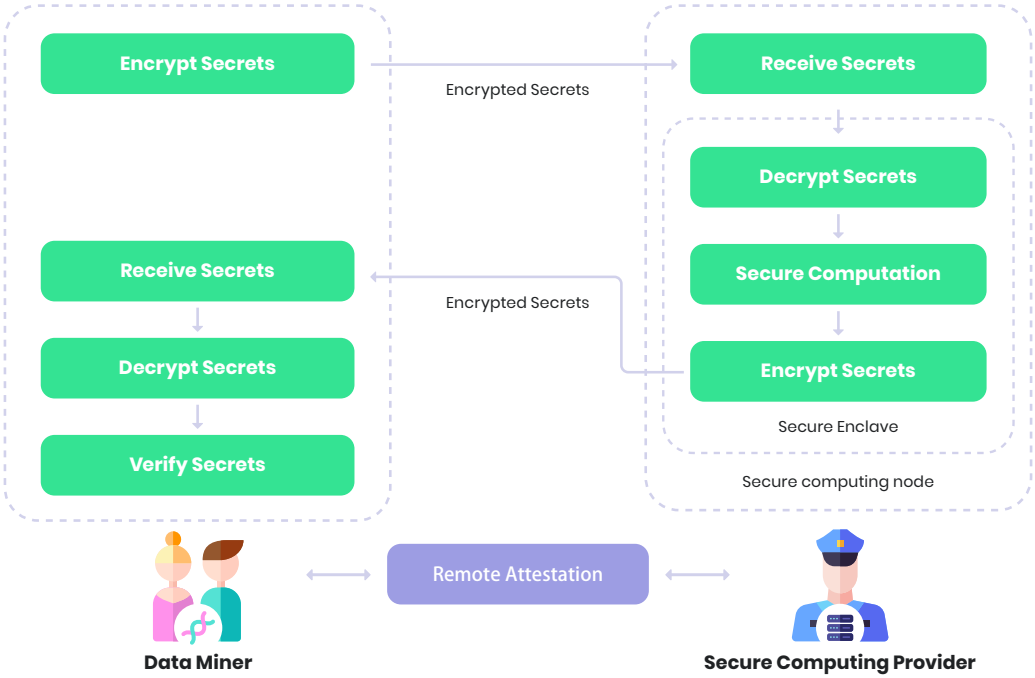


Figure 9: Overview of SGX framework, consisting of data contributors, cloud service provider (CSP), and secure enclaves

Intel SGX-enabled CPU provides the infrastructure for privacy-preserving computation. SGX enclaves are isolated areas of memory called EPC (Enclave Page Cashe) where sensitive data are protected from malicious attacks. The code and data in EPC are reachable only within the secure enclave or authorized parties. Using the Intel SGX platform firmware and software, the Lavita platform reserves EPC for this safe environment.

When applications run inside an enclave, the CPU instantly encrypts it and stores the key. Since the key is inside the CPU, an attacker even with OS root privilege cannot obtain it by inspecting the system memory. Enclaves are extremely safe environments for working with data. What makes enclaves secure is the automatic hardware encryption. The SGX technology uses the CPU to encrypt the information and store the key inside it. Hence, an external party cannot acquire the key and compromise the data. This means that not even the cloud provider can gain access.

Data contributors in the Lavita platform will send their encrypted genomic and healthcare data to the secure computing infrastructure and obtain a hash value that uniquely identifies the data in the Lavita ecosystem. During the secure computing phase, Theta EdgeStore can work directly with Secure Computing Nodes through the identical hash value which is recorded on the blockchain by the data contributor through a smart contract.

In the future, Lavita will evaluate additional distributed computation techniques that augment TEE infrastructure including multi-party computation (MPC) [25, 26, 27, 28, 29, 30], homomorphic encryption (HE) [31, 32, 33, 34], differential privacy (DP) [35, 36, 37, 38, 39, 40, 41, 42], and zero-knowledge proof (ZKP), as described below:

Multi-Party Computation (MPC) represents a collection of data privacy interactive protocols (e.g. secret sharing, garbled circuit, GMW, etc) for computing some functions (represented as circuits) between multiple parties. MPC protocols have the advantage of being theoretically secure rather than relying on computational assumptions as long as there is no collusion or even collusion ratio is under a given limit.

Homomorphic Encryption (HE) allows functions to be computed on ciphertexts without decryption. Technically, any function that can be represented by a low-degree polynomial with addition, subtraction, and multiplication can be supported by HE.

Differential Privacy (DP) offers theoretically quantifiable bounds of privacy on the disclosure of sensitive data or results. A majority of the research work revolving around DP and genomic data lies in the private dissemination of genome-wide association studies (GWAS) and Genomic Beacon Services [43].

Fusion of Secure Computation Technologies. Each secure computation technology has its unique advantages and disadvantages. DP is exceptional for protecting an individual's private contributions within a dataset, but the added noises often prevent fine-grained data analytics on personalized data. MPC and HE can perform addition on encrypted data efficiently, but performing multiplications is very expensive in terms of communication and computation overhead. TEE requires trusting the correctness of the enclave code and the execution environments, but recently discovered vulnerabilities had demonstrated potential risks that leak sensitive information.

Over time, the optimal strategy is to fuse these secure computation technologies together to mitigate the security and performance drawbacks of each technology. For example, TEE technologies may suffer side-channel attacks [44], whereas HE requires a time-consuming noise reduction step after every multiplication over ciphertexts. A longer term strategy is to develop a hybrid solution to combine these two technologies such that straightforward computations can be performed homomorphically within the protection of HE while the computation-intensive noise reduction step is performed within TEE. Lavita's further development plans include solutions with MPC protocols to achieve information-theoretic security for protecting user-sensitive data.

2.3.4 Data Research and Validation

1. **AI-Supported Data Research for Clinical Guidelines**
   Users will receive well-informed genetic information and analysis based on their clinical needs in the platform. Data research applications will be supported on different types of data including genomic, survey, clinical/EHR, and behavioral/longitudinal data.

   With the use of a Semantic Search Engine powered by NLP and NLU, machines are able to detect language patterns and identify relationships between words to understand what people are researching for. Understanding a searcher's intent and the meaning of words and phrases in context to find the right content is the purpose of our Semantic Search Engine, in which collected healthcare data (e.g., symptoms of diseases, clinical guidelines that users may follow for urgent cases, etc.) will be leveraged on the Lavita platform, as shown in Figure 10.

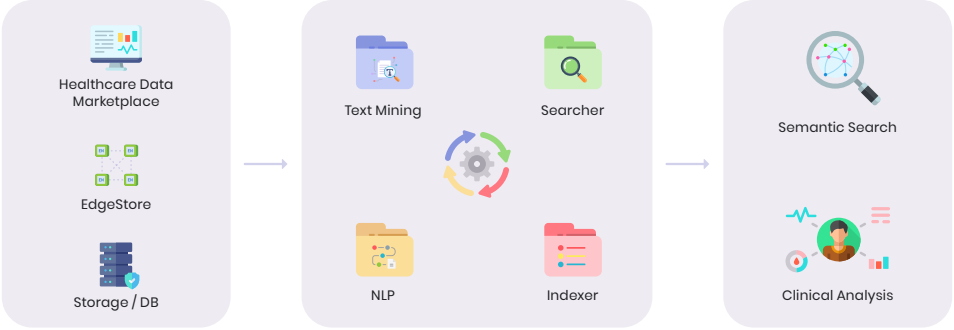


Figure 10: Semantic Search Engine overview

2. **AI-Powered Data Validation Tool**
   For the marketplace to be as competitive in the healthcare industry, it is crucial to have a certain level of data quality. As the platform collects more and more data on the storage infrastructure, analytics capabilities will get more sophisticated, therefore it is crucial to focus more on data quality management for the platform. As the first step to maintain the accuracy and own value of the project, four key dimensions of data quality assessment are defined as follows: 1) Completeness, 2) Consistency, 3) Conformity, 4) Accuracy.

   - **Data Completeness** refers to the meaning that datasets have no or a limited number of missing values. In order to create values with quality data for individuals or healthcare-related organizations, datasets should have all the necessary information to effectively explore and share clinical analysis or guidelines with a certain level of confidence, which is an important point for the Lavita platform.
   - **Consistency** means data across all systems reflects the same information, so the data has to be up to date periodically, to avoid offering faulty information to users.
   - **Conformity** is where data needs to be in the same type, format, etc., to make data more reliable, and easily accessible.
   - **Accuracy** refers to the degree that information accurately reflects a fact or actual data.

   In order to add more liquidity to the Lavita marketplace, one of the incentive mechanisms is with the data quality to make sure that the more efforts individuals or data contributors put into filling up the data survey or results, the better rewards they receive. It is enabled by the AI-Powered Data Quality Tool, as illustrated in Figure 11.

Figure 11: Data quality tools ensure data consistency, conformity, and completeness

2.3.5 Secure Key Management

In Lavita, a trusted execution environment (TEE) is used in the Secure Key Management Services (secure KMS module to protect the security of all keys. Through per-application isolation based on hardware supported encryption, the attack surfaces of the applications using in TEE are significantly reduced. Clients (e.g., data contributors or data miners) can use TEE enabled secure secure KMS to generate dynamic session keys to establish a secure channel in order to deliver their encryption keys to the secure KMS database for future use (see Figure 12).

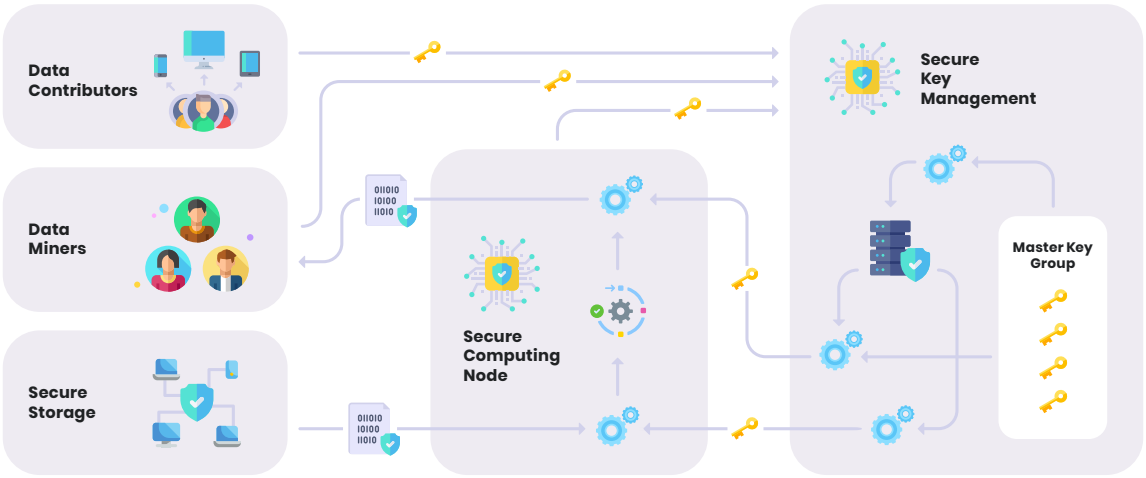


Figure 12: Overview of the secure Key Management Services

Secure KMS provides key hierarchy management to reduce the risks of key breaches. The key hierarchy design allows different encryption keys to be assigned to different data. In the worst case, the breach of a subset of the data will not result in a broad information leakage for the data encrypted by the other keys. In this way, a higher level of security and flexibility is achieved for the protection of keys and their corresponding encrypted data.

3 Token Utility

In order to increase the rate of user adoption in the Lavita ecosystem, the LAVITA token will be created to incentivize data contribution, computation and storage as well as serve as the governance token for the Lavita platform.

In Section 3.1, Figure 13 illustrates the token utility in the Lavita platform of the Data Contributors' side. Data Contributors will earn LAVITA when they upload their own health data powered by smart contracts running on a customized Theta Subchain, and edge nodes providing storage and computation will also be rewarded.

3.1 Utility for Data Contributors

Data Contribution (①–②):
① Data Contributors contribute their data, private and on their device

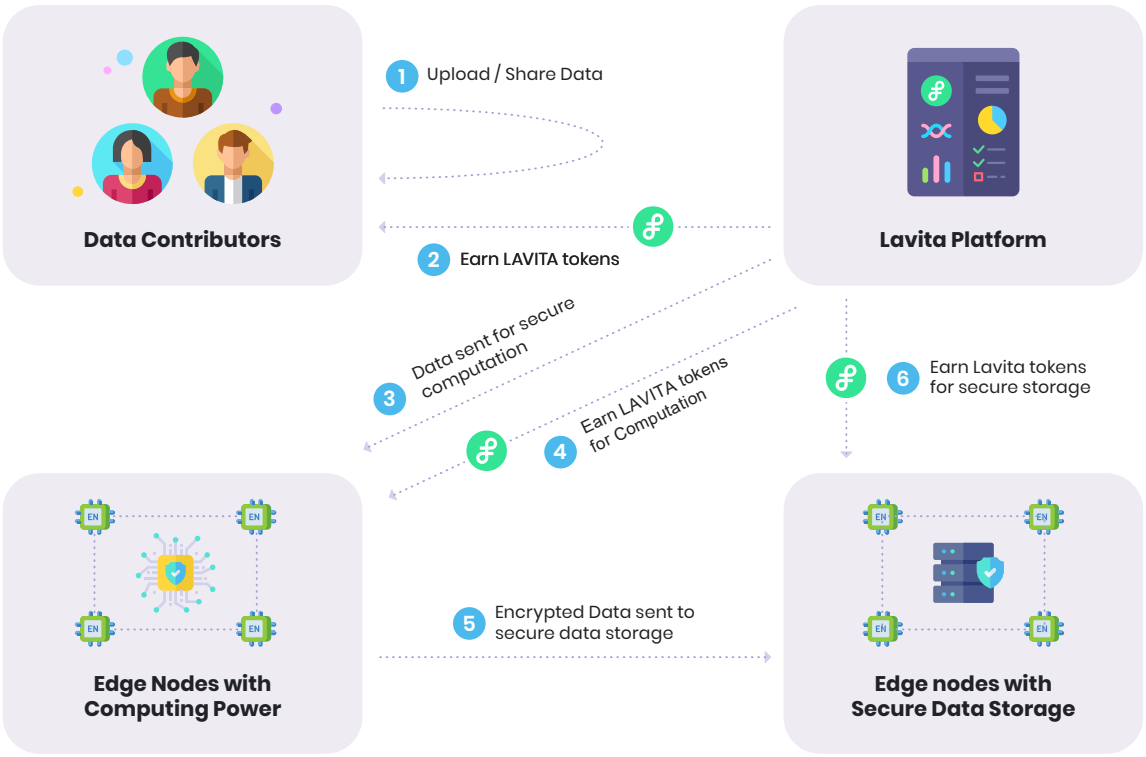


Figure 13: Token utility of the Lavita marketplace (Data Contributors)

② Earn LAVITA tokens with two incentive mechanisms for measuring rewards as below:

- The concept of the 'fair market pricing' algorithm will identify the value of data such as uniqueness, and the amount of tokens will be calculated to reflect the value of data fairly.
- The quality of data is also critical for the platform in order to make the marketplace more reliable and competitive than other healthcare platforms. The amount of tokens will be calculated based on the quality of data by an AI-supported data validation tool.

Secure Data Computation (③–⑤):
③ Once the platform receives data from data contributors, it directly sends data to server-based SGX nodes, for secure computation in an encrypted format.
④ In return, the nodes will be rewarded with LAVITA tokens.
⑤ Once the secure computation is processed, the encrypted data will be sent to secure data storage provided by Edge Nodes.

Secure Storage for Data Results and Analysis (⑥):
⑥ Data results will be stored in Secure Data Storage, and Edge Node providers will be incentivized with LAVITA tokens as rewards.

In Section 3.2, Figure 14 shows the token utility in the Lavita platform of the Data Miners' side. It illustrates that Data Miners need to use LAVITA tokens in order to get access to private healthcare data for results and analysis.

3.2 Utility for Data Miners

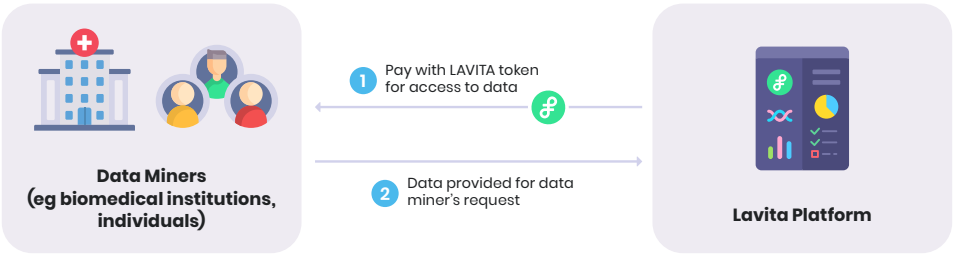


Figure 14: Token utility of the Lavita marketplace (Data Miners)

Data Exchange (①–②):
①–② Data Miners that need access to private genomic or health related data from the Lavita platform will purchase and use LAVITA tokens to gain access to data, analysis, and results.

3.3 Fair Market Pricing Algorithm

Fairness is a core principle of the Lavita marketplace. By giving the control back to the original owners of the assets (data contributors, data verifiers, and data collectors), the platform allows every stakeholder to capture the true value of its asset and thereby maximize the economic surplus. The value of each asset is determined in a fair and decentralized protocol as follows:

Mutual Entropy: The study of rare genetic diseases will be used as an example to describe the mechanism for valuing genomic data. It is important to distinguish between wide-scale genome studies such as Genome-Wide Association Studies (GWAS) for a large population and common diseases and genomic studies for rare diseases. GWAS are observational studies of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. When the population is large and the target trait is observed in a large subset of the population, GWAS relies on a large number of participants to infer statistically significant associations. Thus, the addition or deletion of any specific individual from the population may not dramatically change the outcome of GWAS. On the other hand, when the study is about a rare disease, there are very few individuals that may have the rare condition and thus adding or removing any of those individuals would significantly affect the statistical power of the study. Because of this dynamic, data from an individual with a rare disease is significantly more useful and valuable than a normal individual.

A mathematical method known as mutual entropy is used to evaluate the incremental value that an individual's genome would add to the population if their genome information is included in the study. The entropy of a random variable is a function that characterizes the unpredictability of the random variable. For example, consider a random variable X representing the number that comes up on a roulette wheel and a random variable Y representing the number that comes up on a fair six-sided die. The entropy of X is greater than the entropy of Y because X can take values from numbers 1 through 36, but Y can only take values from 1 through 6. X is essentially less predictable than Y. If a random variable X takes on values in a set $X = \{x_1, x_2, ..., x_n\}$, and is defined by a probability distribution $P(x)$, then the entropy of the random variable is defined as:

$$H(X) = -\sum_{x \epsilon X} P(x) \log P(x)$$

The entropy equation can be used to define Mutual Information between two variables. Mutual information is a measure of the additional information known about one expression pattern when given another. For example, the Mutual Information between variables A and B is defined as:

$$I(A, B) = H(A) - H(A|B)$$

where $H(A)$ is the entropy of the variable A and $H(A|B)$ is the conditional entropy of A given B. Mutual information can be calculated by subtracting the entropy of the joint gene expression patterns from the individual gene entropies.

$$I(A, B) = H(A) + H(B) - H(A, B)$$

Mutual information of zero means that the joint distribution of expression values holds no more information than each of the individual sets considered separately. Higher mutual information between two genes means that one gene is non-randomly associated with the other. Using Mutual Information, any new user that submits their genomic and health information to the platform can fairly price the value of their information, as related to the previously available information on the platform. For example, if a study has already 499 participants, the value of the data that is contributed by the 500th participant is calculated by the mutual entropy between the new genome and each of the first 499 genomes and using the average value as the measure of the incremental value. The new participants will receive higher rewards if their genomic data has higher incremental value.

4 Lavita Open Platform with AI-Powered Dapps

The Lavita platform is an open AI infrastructure that powers distributed applications and interactions among various stakeholders in the healthcare field and creates numerous opportunities for organizations and individuals through health data management, exchange, and utilization. Some of the Lavita platform's initial applications are shown in Figure 15.

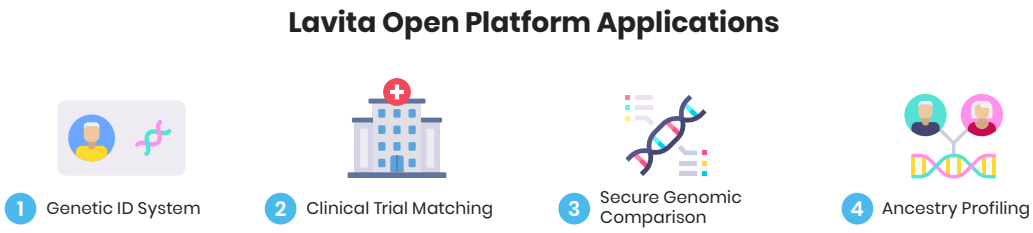


Figure 15: Some initial applications being enabled by Lavita platform

With the help of massive human genomic and healthcare data mining, a much broader opportunity becomes available to biomedical and pharmaceutical researchers to develop targeted drugs and devise novel precision medicine strategies. Lavita's vision is to contribute to the promising clinical research field with its open platform applications and support all kinds of data mining techniques, including those based on machine learning and deep learning algorithms. Equipped with a rich set of data, open-sourced AI algorithms could revolutionize biomedical discovery in the near future. Lavita hopes to play a critical role in this process from data synthesis to new discoveries.

On Lavita platform data contributors will be able to manage the sharing of their genomic and clinical data without worrying about the potential for fraud or a privacy breach and financially benefit from contributing their data to scientific research through a reward system. In addition, customers can enjoy a healthier lifestyle with precision healthcare for disease prevention and medical intervention based on on-chain services for medical care and medical insurance claims.

4.1 Genetic Identity

Genetic identity (Genetic ID) can provide basis for a universal identity system to enable users securely and privately access various health services and applications in Lavita marketplace and elsewhere. It is important to note the Genetic ID system neither stores nor reveals an individual's genomic information, but rather creates a unique and irreversible hash based on user's genomic profile. Such an ID system will not reveal any personal information or genomic data about the individual but provides unique and secure primitives for accessing various applications on the platform. Lavita Genetic ID system will enable a multi-tier user experience (based on user data contributions), giving users exclusive access to digital assets on Lavita platform and unlock other benefits.

Recently, some hospitals have started using DNA forensics, i.e., using short tandem repeats (STR) to identify patients [45]. Similarly, Lavita platform will allow users to encrypt STR information to generate hash as unique ID values. Genetic identity is stable with the usage of each individual's inherent information, and cannot be duplicated or reverse-engineered. Such features make encrypted genetic identity a suitable choice in secure ID systems.

4.2 Clinical Trial Matching

Considering that the average cost of developing and bringing a novel pharmaceutical agent to market comes in around $2.6 billion, and the time invested into a single project can be up to 12 years in total, pharmaceutical companies are increasingly looking for ways to lower costs and maximize returns on their investments.

Now, with a request of the patient query on Lavita secure data marketplace will bring companies a cohort of patients who are the targeted clinical trial participants. Lavita addresses the security and regulatory concerns of both the patients and the medical organizations during the processes of therapeutic intervention and pharmacogenetic clinical trials; the blockchain-enabled Lavita platform is able to reconnect patients and only disclose their data to the company with their complete consent.

Entering virtual clinical trials [46], where participants are not required to travel to a clinical center or doctor's office for frequent recurring inspection, will also be big news for patients, especially those who are suffering from chronic diseases. Lavita blockchain-enabled virtual clinical trials can be a new method of collecting safety and efficacy data from participants.

4.3 Secure Genomic Comparison

This genomic application provides an intuitive and secure app to compare individuals based on their genomic profiles, and in future version based on their health profile as well. The app, partially developed in collaboration with Lavita advisors [47], calculates the similarity of genomics sequences using Hamming distance measure [48] and shows the distribution of phenotypes of similar individuals given the input genomic sequence of the user. The app utilizes publicly available datasets such as Personal Genome Project [49] to amass individuals' profiles along with their genome data, medical conditions, and treatment. Secure Genomic Comparison app will assist biomedical researchers to conduct statistical studies on the human genome, as well as individuals to learn about their own genetics and health. By analyzing your genome data and comparing it to various databases, users can see how certain diseases are common among patients with the most similar genomes to theirs.

In the first version of the Secure Genomic Comparison app, the numbers next to a disease refer to the occurrence of each disease among those people in the database with a similar genomic profile to the user. This information can help you gain a better understanding of a genome profile and its implication for phenotypic traits and health risks. In future versions of the app, the users will be able to choose among different genomic distance measurements besides Hamming distance to quantify the similarity of their genomes to the database. In addition, the users will be able to use health data (i.e., phenotypic data) and genomic variant information to find exact matches among individual profiles in the database.

4.4 Ancestry Profiling

Companies including 23andMe and Ancestry.com have popularized Direct-to-Consumer (DTC) genomic profiling in the last two decades. With the rapid decrease in the cost of genomic sequencing, more people have been able to gain access to sequencing technology through DTC companies and learn from their genomic profile. One of the most common applications popularized by DTC genomic companies is ancestry profiling which provides valuable insights to individuals to understand historical traits and ancestral roots based on their genome. Genetic ancestry is a measure of individuals' biogeographical origins, based on correlated allele frequency differences among ancestral source populations. Knowing genetic ancestry can provide useful insights about the geographic origins of individuals' ancestors and even aid in the assessment of risk for some heritable conditions.

In this application, by leveraging genome-wide genetic variant data, including whole-genome sequences, whole exome sequences, and whole-genome genotypes, we analyze users' raw genome data reports (e.g., raw reports obtained from 23andMe or other genomic databases) and determine their genetic ancestry categories. In particular, by leveraging our bioinformatic and machine learning tools and pipelines, we find and visualize the fractional estimates of ancestry components for genomic samples of users in a secure environment. In future versions of the app, the users will be able to obtain more granular geographic information about their ancestral relatives and be able to opt-in to securely connect with other individuals with similar ancestral profiles.

5 New Business Opportunities

5.1 Pharmacy Benefit Management (PBM)

Pharmacies participating in the Lavita ecosystem can continue to increase the precision applied to their policies and clinical programs based on what the most recent pharmacology research offers. Pharmacy companies are able to improve drug traceability, simplify the process of interaction between parties in the supply chain, and even alert labs and pharmacy outlets if fake drugs have been detected [50].

5.2 On-chain Medical Insurance and Medical Care

Let's take a look at an individual who purchases health insurance. The details of his/her policy get linked to the corresponding profile within the blockchain. When a patient undergoes a medical procedure covered by his/her policy, a smart contract will automatically be triggered and the correct payment from the insurance company to the hospital will be made [51]. This will have a positive impact on the fair execution of a person's insurance policy. It will reduce the inefficiencies to complete insurance claims forms. On the other hand, a company can retrieve data and make sure the documents are not fraudulent.

6 LAVITA Token Metrics

LAVITA tokens will have a fixed supply of 8,000,000,000 (8 billion) tokens with a tentative launch of May 1, 2023, subject to change. The token allocation will be as follows:

Tokens for contributing to Lavita platform
25% - Rewards for Data Contributors on Lavita platform (4-year period)
15% - Rewards for LAVITA Staking and Decentralized Governance (4-year period)

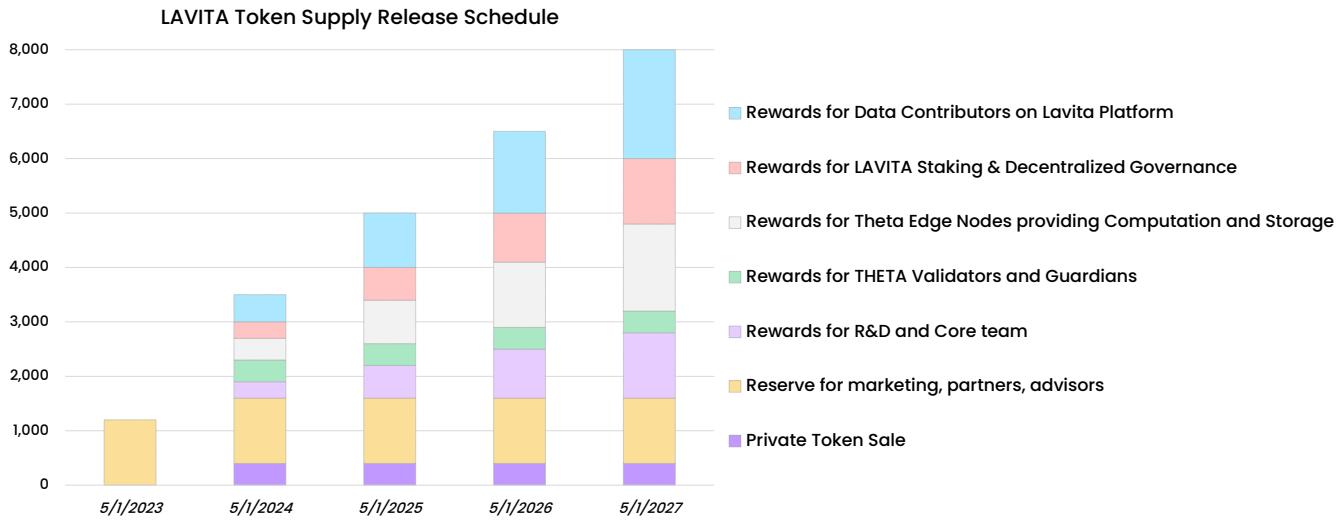
Tokens for Theta ecosystem partners
20% - Rewards for Theta Edge Nodes providing Computation and Storage (4-year period)
5% - Rewards for THETA Validators and Guardians (12-month vesting, 1/4th per quarter)

Tokens for Lavita team/marketing & private sale
15% - Rewards for R&D and Core team (4-year vesting, 1/4th per year)
15% - Reserve for marketing, partners, advisors (not vested)
5% - Private Token Sale (12-month vesting, 1/4th per quarter)

The expected amount of LAVITA in circulation is shown in the following graph and detailed in the table below(# of tokens in Millions).

Rewards for Data Contributors: A total of 2 Billion LAVITA (25% of supply) is allocated over a 4-year period for data contributors who provide their genomic-health data on the Lavita

Token Allocations (# of tokens in Millions)	%	Total	5/1/2023	8/1/2023	11/1/2023	2/1/2024	End of Yr 1 5/1/2024	End of Yr 2 5/1/2025	End of Yr 3 5/1/2026	End of Yr 4 5/1/2027
Rewards for Data Contributors on Lavita platform (4-year period)	25%	2,000	0	125	250	375	500	1,000	1,500	2,000
Rewards for LAVITA Staking & Decentralized Governance (4-year period)	15%	1,200	0	75	150	225	300	600	900	1,200
Rewards for Theta Edge Nodes providing Computation and Storage (4-year period)	20%	1,600	0	100	200	300	400	800	1,200	1,600
Rewards for THETA Validators and Guardians (12-month vesting)	5%	400	0	100	200	300	400	400	400	400
Rewards for R&D and Core team (4 year vesting, 1/4th per year)	15%	1,200	0	0	0	0	300	600	900	1,200
Reserve for marketing, partners, advisors (Not vested)	15%	1,200	1,200	1,200	1,200	1,200	1,200	1,200	1,200	1,200
Private Token Sale (12-month vesting, 1/4th per quarter)	5%	400	0	100	200	300	400	400	400	400
Total Circulating	100%	8,000	1,200	1,700	2,200	2,700	3,500	5,000	6,500	8,000
Assume 50% circ staked (APY %)			*50.00%*	*35.29%*	*27.27%*	*22.22%*	*17.14%*	*12.00%*	*9.23%*	*7.50%*
Amount of tokens staked (est 50%)			600	850	1100	1350	1750	2500	3250	4000



marketplace. The "fair market pricing" algorithm and AI-supported data validation tool will identify the value of data and the amount of tokens to be distributed fairly.

Rewards for LAVITA Staking & Decentralized Governance: A total of 1.2 Billion LAVITA (15% of supply) is allocated over a 4-year period as rewards for staking LAVITA, which also allows users to participate in the decentralized governance of the network. Each year, 300 Million LAVITA will be distributed proportionally to all LAVITA stakers. At the end of the 4-year period, the community can propose and vote on possibly extending this.

Rewards for Theta Edge Nodes providing Computation and Storage: For edge nodes supporting Lavita computation and storage, a total of 1.6 Billion LAVITA (20% of supply) is allocated over a 4-year period.

Rewards for THETA validators and guardians: on August 15, 2023, 100 Million LAVITA (or $\frac{1}{4}$ of the allocated 400 Million) will be distributed proportionally to all THETA validator and guardian stakers. An average amount of THETA staked between May 1, 2023 and July 31, 2023 will be used for reward computation. For example, if a guardian stakes an average 100,000 THETA during those 3 months, then they will receive 100,000 / total ave THETA staked x 100 Million LAVITA.

Thereafter, 100 Million LAVITA (or $\frac{1}{4}$ of the allocated 400 Million LAVITA) will be distributed

proportionally to stakers each quarter through May 1, 2024.

Rewards for R&D and Core team: A total of 1.2 Billion LAVITA (15% of total supply) is allocated to R&D and core team to incentivise continued development of the platform. On May 1, 2024, 300 Million LAVITA will become vested and distributed to the team and 300 Million LAVITA each year thereafter through May 1, 2027.

Reserve for marketing, partners, advisors: A total of 1.2 Billion LAVITA (15% of total supply) is reserved for marketing purposes, advisors, partners and other strategic development purposes. These tokens are not vested.

Private Token Sale: A total of 400 Million LAVITA (5% of total supply) is allocated for private sale. If all 400 Million tokens are sold prior to May 1, 2023 then on August 1, 2023, 100 Million LAVITA will become vested, and 100 Million LAVITA will be distributed each quarter thereafter through May 1, 2024.

7 Conclusion

The long-term vision for the Lavita genomic and healthcare data marketplace is to create an environment for users to actively participate and be rewarded with LAVITA tokens for sharing their health data and offering storage and computation capabilities on the Theta Edge Network. In the end, we hope to create a new global community for customer adoption and accelerate the growth of the Lavita ecosystem.

In addition to the incentive mechanism, users will be provided with AI-supported personalized clinical advice and health data analysis. This drives further engagement within the Lavita platform and increased usage among patients for clinical guidelines. With the recent advances in conversational and generative AI technologies and distributed ledger and blockchain technologies, Lavita aims to revolutionize biomedical discoveries leading to novel, more effective healthcare diagnostic and therapeutic tools.

The next phase in Lavita's development will be to create new business opportunities in the following areas:

1. Clinical Patient Matching including possibly virtual clinical trials that could lead to significant cost and time savings for pharmaceutical companies to invest in developing novel drugs and treatments,
2. Pharmacy Benefit Management systems to improve drug quality and traceability, and
3. On-chain Medical Insurance programs leveraging smart contracts to automatically process claims and payments.

References

[1] Gartner forecasts worldwide public cloud end-user spending to reach nearly $600 billion in 2023. https://www.gartner.com/en/newsroom/press-releases/2022-10-31-

gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-nearly-600-billion-in-2023.

[2] Nih security best practices for controlled-access data subject to the nih genomic data sharing (gds) policy. https://sharing.nih.gov/sites/default/files/flmngr/NIH_Best_Practices_for_Controlled-Access_Data_Subject_to_the_NIH_GDS_Policy.pdf.

[3] The cost of sequencing a human genome," national human genome research institute (nhgri). http://www.genome.gov/sequencingcosts/.

[4] Lincoln D Stein. The case for cloud computing in genome informatics. *Genome biology*, 11(5):1–7, 2010.

[5] A Regalado. 2017 was the year consumer dna testing blew up. mit technology review. 2018.

[6] Antonio Regalado. More than 26 million people have taken an at-home ancestry test. *MIT Technology Review*, 11(2):2019, 2019.

[7] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[8] R Shouval, O Bondi, H Mishan, A Shimoni, R Unger, and Arnon Nagler. Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct. *Bone marrow transplantation*, 49(3):332–337, 2014.

[9] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Feng Chen, Michelle Dow, Sijie Ding, Yao Lu, Xiaoqian Jiang, Hua Tang, and Shuang Wang. Premix: Privacy-preserving estimation of individual admixture. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1747. American Medical Informatics Association, 2016.

[12] Feng Chen, Chenghong Wang, Wenrui Dai, Xiaoqian Jiang, Noman Mohammed, Md Momin Al Aziz, Md Nazmus Sadat, Cenk Sahinalp, Kristin Lauter, and Shuang Wang. Presage: privacy-preserving genetic testing via software guard extension. *BMC medical genomics*, 10(2):77–85, 2017.

[13] Feng Chen, Shuang Wang, Xiaoqian Jiang, Sijie Ding, Yao Lu, Jihoon Kim, S Cenk Sahinalp, Chisato Shimizu, Jane C Burns, Victoria J Wright, et al. Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. *Bioinformatics*, 33(6):871–878, 2017.

[14] Wenhao Wang, Guoxing Chen, Xiaorui Pan, Yinqian Zhang, XiaoFeng Wang, Vincent Bindschaedler, Haixu Tang, and Carl A Gunter. Leaky cauldron on the dark land: Understanding memory side-channel hazards in sgx. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2421–2434, 2017.

[15] Karthik A Jagadeesh, David J Wu, Johannes A Birgmeier, Dan Boneh, and Gill Bejerano. Deriving genomic diagnoses without revealing patient genomes. *Science*, 357(6352):692–695, 2017.

[16] Hyunghoon Cho, David J Wu, and Bonnie Berger. Secure genome-wide association analysis using multiparty computation. *Nature biotechnology*, 36(6):547–551, 2018.

[17] Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[18] Oded Goldreich, Silvio Micali, and Avi Wigderson. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 307–328. 2019.

[19] Kim Laine and Rachel Player. Simple encrypted arithmetic library-seal (v2. 0). *Technical report, Technical report*, 2016.

[20] Shai Halevi and Victor Shoup. Algorithms in helib. In *Annual Cryptology Conference*, pages 554–571. Springer, 2014.

[21] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. *ACM Transactions on Computation Theory (TOCT)*, 6(3):1–36, 2014.

[22] Shuang Wang, Yuchen Zhang, Wenrui Dai, Kristin Lauter, Miran Kim, Yuzhe Tang, Hongkai Xiong, and Xiaoqian Jiang. Healer: homomorphic computation of exact logistic regression for secure rare disease variants analysis in gwas. *Bioinformatics*, 32(2):211–218, 2016.

[23] Eiichiro Fujisaki and Tatsuaki Okamoto. Statistical zero knowledge protocols to prove modular polynomial relations. In *Annual International Cryptology Conference*, pages 16–30. Springer, 1997.

[24] Intel® software guard extensions (intel® sgx). `https://software.intel.com/en-us/isa-extensions/intel-sgx`.

[25] Somesh Jha, Louis Kruger, and Vitaly Shmatikov. Towards practical privacy for genomic computation. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 216–230. IEEE, 2008.

[26] Xiao Shaun Wang, Yan Huang, Yongan Zhao, Haixu Tang, XiaoFeng Wang, and Diyue Bu. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 492–503, 2015.

[27] Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K Reiter, and Zheng Dong. Privacy-preserving genomic computation through program specialization. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 338–347, 2009.

[28] Gilad Asharov, Shai Halevi, Yehuda Lindell, and Tal Rabin. Privacy-preserving search of similar patients in genomic data. *Cryptology ePrint Archive*, 2017.

[29] Ruiyu Zhu and Yan Huang. Efficient privacy-preserving general edit distance and beyond. *Cryptology ePrint Archive*, 2017.

[30] Md Momin Al Aziz, Dima Alhadidi, and Noman Mohammed. Secure approximation of edit distance on genomic data. *BMC medical genomics*, 10(2):55–67, 2017.

[31] Zihao Shan, Kui Ren, Marina Blanton, and Cong Wang. Practical secure computation outsourcing: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–40, 2018.

[32] Thomas A Hemphill and Phil Longstreet. Financial data breaches in the us retail economy: Restoring confidence in information technology security standards. *Technology in Society*, 44:30–38, 2016.

[33] Tyler Moore. On the harms arising from the equifax data breach of 2017. *International Journal of Critical Infrastructure Protection*, 19(C):47–48, 2017.

[34] Nathan Manworren, Joshua Letwat, and Olivia Daily. Why you should care about the target data breach. *Business Horizons*, 59(3):257–266, 2016.

[35] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087, 2013.

[36] Fei Yu, Michal Rybar, Caroline Uhler, and Stephen E Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014.

[37] Fei Yu and Zhanglong Ji. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to idash healthcare privacy protection challenge. *BMC medical informatics and decision making*, 14(1):1–8, 2014.

[38] Florian Tramèr, Zhicong Huang, Jean-Pierre Hubaux, and Erman Ayday. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1286–1297, 2015.

[39] Sean Simmons and Bonnie Berger. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9):1293–1300, 2016.

[40] Sean Simmons, Cenk Sahinalp, and Bonnie Berger. Enabling privacy-preserving gwass in heterogeneous human populations. *Cell systems*, 3(1):54–61, 2016.

[41] Caroline Uhlerop, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5(1):137, 2013.

[42] Shuang Wang, Noman Mohammed, and Rui Chen. Differentially private genome data dissemination through top-down specialization. *BMC medical informatics and decision making*, 14(1):1–7, 2014.

[43] Md Momin Al Aziz, Reza Ghasemi, Md Waliullah, and Noman Mohammed. Aftermath of bustamante attack on genomic beacon service. *BMC medical genomics*, 10(2):43–54, 2017.

[44] Shuo Chen, Rui Wang, XiaoFeng Wang, and Kehuan Zhang. Side-channel leaks in web applications: A reality today, a challenge tomorrow. In *2010 IEEE Symposium on Security and Privacy*, pages 191–206. IEEE, 2010.

[45] Greggory S LaBerge, Eric Duvall, Zachary Grasmick, Kay Haedicke, and John Pawelek. A melanoma lymph node metastasis with a donor-patient hybrid genome following bone marrow transplantation: A second case of leucocyte-tumor cell hybridization in cancer metastasis. *PloS one*, 12(2):e0168581, 2017.

[46] Sujay Jadhav. Virtual clinical trials: the future of patient engagement. *Applied Clinical Trials*, 2016.

[47] Lichang Wang, Yong Fang, Dima Aref, Suyash Rathi, Li Shen, Xiaoqian Jiang, and Shuang Wang. Palme: Patients like my genome. *AMIA Summits on Translational Science Proceedings*, 2016:219, 2016.

[48] Hildete Prisco Pinheiro, Aluísio de Souza Pinheiro, and Pranab Kumar Sen. Comparison of genomic sequences using the hamming distance. *Journal of Statistical Planning and Inference*, 130(1-2):325–339, 2005.

[49] Bryony Jones. Personal genome project. *Nature Reviews Genetics*, 13(9):599–599, 2012.

[50] J Russell Teagarden and Eric J Stanek. On pharmacogenomics in pharmacy benefit management. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 32(2):103–111, 2012.

[51] David Randall, Pradeep Goel, Ramzi Abujamra, et al. Blockchain applications and use cases in health information technology. *Journal of Health & Medical Informatics*, 8(3):8–11, 2017.

[52] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[53] Cheryl Erwin. Behind the genetic information nondiscrimination act of 2008. *AJN The American Journal of Nursing*, 109(12):46–48, 2009.

[54] Privacy in genomics. `https://www.genome.gov/about-genomics/policy-issues/Privacy`.

[55] A Gutmann, J Wagner, Y Ali, AL Allen, JD Arras, BF Atkinson, and RS Kucherlapati. Privacy and progress in whole genome sequencing. *Presidential Committee for the Study of Bioethical*, (2012), 2012.

[56] Heather Kelly. California passes strictest online privacy law in the country, 2019.

[57] Premier Li Peng. Ordinance of the people's republic of china on the protection of computer information system security: Decree of the state council of the people's republic of china (no. 147). the" ordinance of the people's republic of china on the protection of computer information system security" is to be implemented as of the day of issuance. february 18, 1994. *Chinese Law & Government*, 43(5):12–16, 2010.

[58] Richard Bird and PY Quah. Where are we now with data protection law in china. *Freshfields-Bruckhaus Deringer Report*, 2019.

[59] Sara Xia. China's personal information security specification: Get ready for may 1. *China Law Blog*, 28, 2018.

[60] Samantha Beaumont. The data protection directive versus the gdpr: Understanding key changes, 2018.

[61] International privacy standards. https://www.eff.org/issues/international-privacy-standards.

Appendix

A Compliance with National and State Laws

Individual's health information privacy is a ubiquitous problem around the world. Activities involving health information face significant challenges, risks, as well as regulations related to individual's privacy.

In the United States, Health Insurance Portability and Accountability Act (HIPAA) [52] regulated how covered entities (e.g., healthcare organizations and their business associates) handle protected health information (PHI). More specifically, the HIPAA privacy rule provides two mechanisms (i.e., safe harbor and expert determination) to de-identify PHI for secondary use. Both methods are based on the consensus that there is "no reasonable basis" for believing that the processed data can be used to re-identify individuals. The former mechanism explicitly defines 18 identifiers to be removed from the data, while the latter mechanism relies on an expert to certify the risk of re-identification. For genetic information, the Genetic Information Nondiscrimination Act (GINA) [53] prevents genetic discrimination in health insurance and in an employment decision, but with the lack of protection against life insurance, disability insurance, or long-term care insurance-related discriminations. Although GINA provides anti-discrimination protection, it is still lacking comprehensive privacy protections of genetic information. As GINA clarified that genetic information is health information in 2013, the HIPAA privacy rule was revised to cover the protection of genetic information [54]. But, the use or disclosure of de-identified PHI is not restricted under HIPAA. At the state level, different U.S. states may have different medical or genetic privacy-related laws. For example, some state laws protect against the improper collection of genetic material without consent [55]. As another example, California's Confidentiality of Medical Information Act (CMIA) even provides stronger privacy protections for medical information than HIPAA, where the definition of providers of healthcare under CMIA (i.e., any business that designed to maintain medical information) is much broader than that of covered entities under HIPAA. In 2018, California Consumer Privacy Act (CCPA) is passed to provide residents full control over their personal information, such as what information will be collected for which purpose and with whom to be shared. CCPA is pioneering consumer privacy protection in California, which will be a blueprint for other states. CCPA also allows consumers to opt out from having their data sold (but not for "sharing" purposes) by a company, but the company also reserves the right to charge a higher price of services for these consumers [56]. In addition, to ensure the confidentiality and integrity of health information, the HIPAA security rule provides national standards for handling (e.g., storage, transferring, use) health information within covered entities. The research use of health information with human subjects is also regulated by the US Department of Health and Human Services Common Rule, which includes three key elements i.e., requirements of assuring compliance by research institutions, obtaining individual's informed consent, and Institutional Review Board (IRB) approvals.

In China, under the regulation of "Cyber Security Law of the People's Republic of China" [57] in 2017, personal information without proper de-identification or data contributors' consent cannot be shared with any third party by any network providers. The cyber security law defines a 'network' as "any system that consists of computers or other information terminals, and related equipment for collecting, storing, transmitting, exchanging and processing information" [58].

In addition, the Chinese government released a new regulation named "Information Security Tech-

nology and Personal Information Security Specification" (referred as "Specification") [59] in 2017. Several security requirements have been provided under this Specification for handling personal data during the data collection, storage, processing, transferring and disclosure phases. The Specification also defines a few exceptions (e.g., for public safety, public health, major public interests, legal rights, etc.) in which explicit consent from individuals may not be required for sharing sensitive personal information.

In Europe, the Data Protection Directive has been designed to protect personal data privacy as national law by the 27 European Union members since 1995. On May 25, 2018, it was replaced by the General Data Protection Regulation (GDPR). GDPR extended the definition of personal data to reflect changes in technology. Under GDPR, any information that could be used to identify an individual (either on its own or by linking with other external information) is defined as personal data. For example, the collection of users' browser history or purchase history (defined as personal data under GDPR) requires explicit consent from users. GDPR also offers EU residents better control over how their data is used [60]. Processing different types of personal data requires explicit "opt-in" informed consent. Moreover, EU residents have the right to access their personal data from data controllers along with additional information (e.g., how, where, and for what purpose their data is being used) without charge. EU residents can also request their data to be removed from a database or any further use.

Worldwide, Personal Information Protection and Electronic Documents Act (PIPEDA) is the Canadian law for data privacy. South Korea and Japan, both have their own versions of the Personal Information Protection Act (PIPA) and a more complete list of international privacy-related laws by regions can be found at [61].