Cryptographic Open Science: Enabling Secure and Incentivized Biomedical Data Sharing with Web 3.0 Technologies to Overcome the Open Science Dilemma

Yu YoSean Wang[1,2,3,6], Junfeng Fan[2, 3], Zhiwei Bao[2], Liang Liu[2], Mengsu Yang[1,4], George M. Church[5], Zhang Sheng[1]

1. Institute of Digital Medicine, City University of Hong Kong, Hong Kong SAR, China
2. DeSci Sino, Global
3. Open Security Research, Shenzhen, Guangdong, China
4. Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China
5. Department of Genetics, Harvard Medical School, Boston, MA, USA
6. Lead Contact

Correspondence: yosean.wang@cityu.edu.hk; gchurch@genetics.med.harvard.edu

Abstract

Biomedical data underpins scientific discovery, and open science offers the potential to accelerate innovation through the unrestricted sharing of knowledge, methodologies, and datasets. However, the **open science dilemma** persists, as researchers hesitate to share data due to privacy concerns, intellectual property risks, and lack of recognition. Stringent data privacy regulations further compound these challenges, limiting data sharing.

To address these barriers, we propose the **Cryptographic Open Science** (COS) framework, integrating advanced technologies for secure, privacy-preserving, and incentivized data sharing. **Blockchain technology** provides immutable records of data ownership and usage, enhancing transparency and trust, while smart contracts automate access controls and enforce compliance. However, blockchain alone does not prevent loss of control over plaintext data once released. COS incorporates **Fully Homomorphic Encryption** (FHE) to allow computations on encrypted data, ensuring end-to-end confidentiality and maintaining full ownership control. Recognizing that privacy alone does not incentivize sharing, COS introduces a **crypto token-based system** to create a market-driven flywheel.

This system rewards contributors and aligns stakeholder interests, promoting active data sharing. By integrating blockchain, FHE, and token-based incentives, COS bridges the gap between the ideals of open science and the practical concerns of data providers, accelerating progress in fields like precision medicine and genomics.In our implementation, we instantiate COS on AntChain Open Alliance with a non-transferable SBT-anchored access-control contract that records verifiable usage events and issues short-lived permits to drive off-chain execution over encrypted data (including optional FHE), ensuring that no plaintext biomedical data resides on-chain.

Keywords: Web 3.0 technologies, Open Science, Biomedical data, Blockchain, Fully Homomorphic Encryption, Cryptocurrency

Introduction

In today's digital age, especially with the rapid growth of artificial intelligence, data has become a vital resource driving advancements across numerous scientific fields. For instance, the second phase of the Human Genome Project (HGP2) has set an ambitious goal to collect multi-omics data from 1% of the global population—equating to 80 million people out of 8 billion[1]. Open science initiatives, which advocate for the open sharing of scientific knowledge, data, and methodologies, have the potential to revolutionize research by facilitating data exchange[2-4]. Significant scientific breakthroughs, such as the rapid development of COVID-19 vaccines, have been greatly accelerated by open data sharing and collaborative efforts within the global scientific community[5].

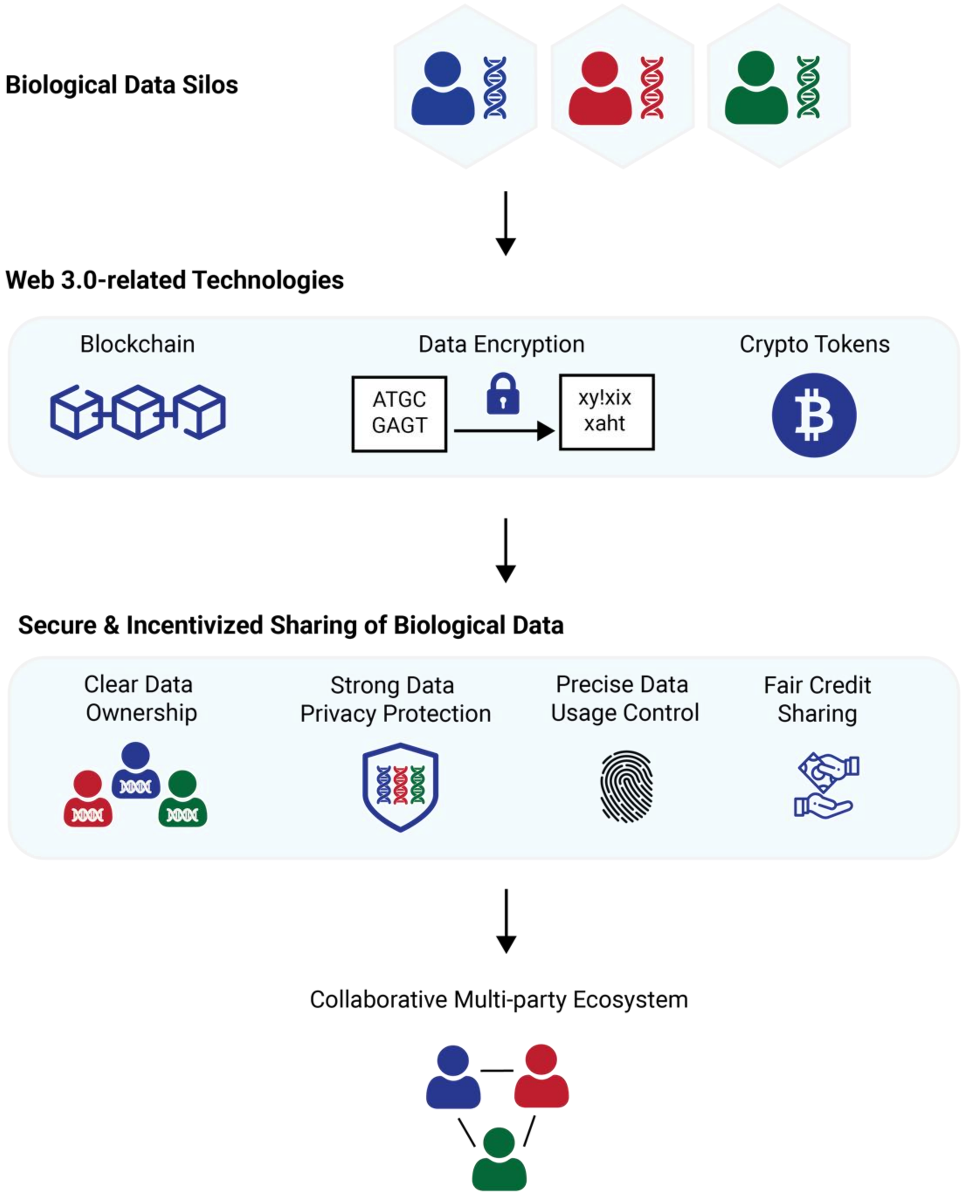

Biological Data Silos
Web 3.0-related Technologies
Blockchain
Data Encryption
ATGC
GAGT
xy!xix
xaht
Crypto Tokens
Secure & Incentivized Sharing of Biological Data
Clear Data Ownership
Strong Data Privacy Protection
Precise Data Usage Control
Fair Credit Sharing
Collaborative Multi-party Ecosystem

Figure 1. Secure and incentivized sharing of biological data using Web 3.0-releated technologies.

Despite the promising benefits of open science, a fully free and transparent system is not always practical[6]. This issue, often referred to as the 'Open Science Dilemma,' is analogous to the classic 'Prisoner's Dilemma' in game theory, where individuals must choose between acting for the collective good or prioritizing personal interests. In the realm of open science, researchers face similar trade-offs. Although open data sharing can greatly benefit the scientific community by fostering faster discovery and wider collaboration, individual researchers often hesitate to participate due to concerns over losing intellectual property rights, recognition, or competitive advantage[6]. Additionally, the absence of sufficient incentives discourages participation, especially when maintaining high data quality requires substantial extra effort. This problem is particularly pronounced in the private sector, where significant volumes of valuable data are held but cannot be shared openly due to IP constraints[7].

In addition, data safety and privacy concerns, particularly for sensitive human-related data, further compound these challenges. Strict data protection laws in regions such as the U.S., Europe, and China—like HIPAA, GDPR and PIPL—create significant legal and ethical barriers to open sharing[8-10]. According to IBM's 2024 Cost of a Data Breach Report, the global average total cost of a data breach is $4.88 million[11]. In 2023, the genetic testing company 23andMe experienced a major hacker attack, compromising the personal information of nearly 7 million customers, including sensitive genetic data. This incident not only led to widespread public concern but also sparked class-action lawsuits against the company[12]. Another notable breach is the 2018 data breach at MyHeritage, affecting 92 million users[13]. These incidents underscore the vulnerability of biological data security.

To address these challenges, we propose a 'Cryptographic Open Science' framework that leverages advanced Web 3.0-related cryptography technologies such as blockchain, data encryption (particularly Fully Homomorphic Encryption, FHE) and crypto tokens to create robust data protection mechanisms that safeguard security, intellectual property, and privacy while incentivizing participation (**Figure 1**). Although it may seem counterintuitive, sharing data with cryptographic protections allows for secure, controlled data exchange without compromising privacy, data ownership, or competitive advantage. This balanced approach between openness and security provides a viable path forward for advancing open science, ensuring that data can be shared in a way that maintains trust and encourages collaboration.

Blockchain Technology: Empowering Data Ownership and Transparent Data Usage

Blockchain is a decentralized, distributed ledger technology that allows data to be recorded, shared, and maintained across a network of computers without the need for a central authority[14,15]. Each block in the blockchain contains a list of transactions or data entries, a timestamp, and a cryptographic link to the previous block, forming a secure and immutable chain. The use of blockchain techniques ensures that once data is recorded, it cannot be altered without the consensus of the network participants (**Figure 2**).

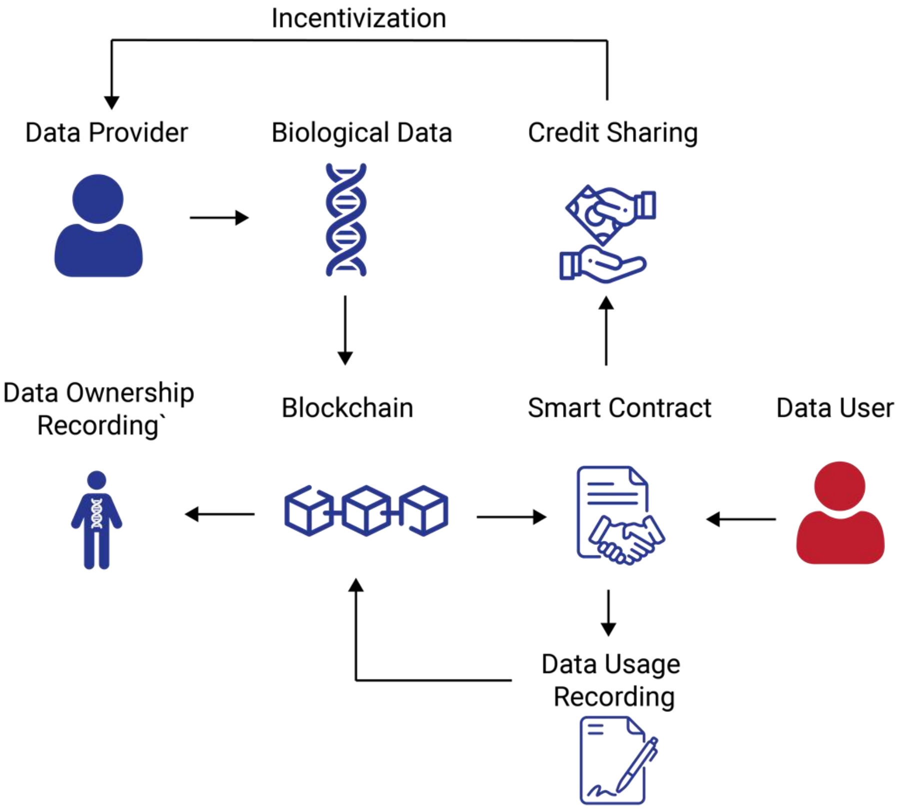


Figure 2. Blockchain technologies used for data ownership recording, smart contract formation between Data Provider and Data User to record data usage and share credits.

In the context of open science, blockchain can be utilized to create immutable records of data, including data provenance, ownership, and usage[16]. These immutable records provide proof of data origin and ownership. By ensuring that original data contributors can be properly traced, blockchain enhances accountability and recognition within the research community. Additionally, the transparent nature of blockchain allows for easy auditing of data collection, access, and usage, enabling data users to verify whether the data resources and collection procedures meet their standards—a factor particularly important when handling sensitive human-related data[17]. A key concept embedded in the Web 3.0 community is "Don't trust, verify," derived from the immutability and transparency of blockchain. Recording data usage on the blockchain enables data providers and owners to track and examine each instance of their data being used. This transforms the traditional collaboration paradigm from building trusted networks to establishing verifiable networks, which is critical for accelerating open science among a large group of scientists.

Researchers can also leverage blockchain-enabled smart contracts, which are self-executing agreements that automatically enforce predefined terms and conditions, to facilitate secure and efficient data sharing[18]. For example, smart contracts can specify detailed access conditions for data sharing by defining user permissions—setting who can access the data, under what conditions, and for how long; authenticating identities by interacting with on-chain identification systems to verify the identity of data requesters, ensuring that only authorized individuals or entities can access sensitive information; and granting conditional access by providing data only when specific criteria are met, such as approval from a data governance board or completion of ethical training modules. By automating access control in this manner, smart contracts reduce the administrative burden on researchers and institutions, eliminating the need for manual verification and approval processes. In addition, smart contracts can facilitate compliance with privacy regulations like GDPR, HIPAA, or other domain-specific laws by embedding legal requirements directly into their clauses to enforce compliance with relevant laws, ensuring data is handled appropriately. They can automatically trigger data anonymization or pseudonymization processes before granting access, enhancing privacy protection. Additionally, smart contracts manage consent by recording and enforcing participant consent parameters, ensuring data is used only in ways that participants have agreed to. This automated compliance helps protect data providers and users from legal risks and enhances the ethical handling of data.

CancerDao Implementation: SBT-Anchored Access Control on AntChain and Its FHE Interface

Platform and design goals. We instantiate the COS design in CancerDao on AntChain Open Alliance, a permissioned consortium blockchain suited for audited biomedical workflows. The ledger stores no plaintext biomedical data; only identifiers (e.g., datasetId), cryptographic commitments, policy hashes, and immutable events are recorded. Our objectives are: (i) verifiable identity binding via non-transferable Soulbound Tokens (SBTs), (ii) least-privilege access control that encodes legal/ethical constraints as explicit on-chain clauses, (iii) a ticketed gateway that enforces grants off-chain against encrypted objects, and (iv) an optional FHE compute path that preserves end-to-end confidentiality while reusing the same permit/audit pipeline.

On-chain artifacts and roles. Identity of a data owner (Owner) is anchored by a non-transferable SBT. The main roles are the Grantee (data user), the Gateway (off-chain access-ticket service), and the Compute Orchestrator (optional, for encrypted/FHE jobs). We implement two contract services: AccessACL (policy, grants, revocation, and audit events) and UsageLog (append-only usage events; not shown in Fig. X for brevity). Figure X(a) summarizes the components and data flow.

SBT-anchored access-control clauses. Each grant is scoped by a concrete, minimal set of clauses that make compliance verifiable on chain:

(1) Identity binding: the grant is attached to sbtId (non-transferable).

(2) Scope: {datasetId/resourceURI, opMask ∈ {READ, LIST, EXPORT, COMPUTE}}.

(3) Purpose limitation: purposeHash commits to the declared purpose (GDPR-style).

(4) Temporal bounds: notBefore, expiry, optional maxRuns/rate-limit.

(5) Residency/geofence: regionCode must pass oracle checks before use.

(6) Privacy budget (optional): e.g., ε/δ caps or per-query counters.

(7) Output policy: by default results remain encrypted for the Owner; key-switching or multi-key FHE is permitted only if authorized by termsHash.

(8) Revocation: the Owner can revoke grants pre-use; post-use follows settlement terms.

(9) Auditability: mandatory state-transition events (Requested, Granted, Used, Revoked).

(10) Settlement hooks (optional): per-use fees and royalty splits upon Used/JobResult.

Baseline access path (no plaintext on chain). The off-chain Gateway issues a short-lived access ticket only when an on-chain grant exists and is valid at issuance time (identity/scope/time/region checks against AccessACL). The minimal on-chain check and ticket issuance logic is shown in Listing S1. Encrypted storage serves objects only upon presentation of a valid ticket (as depicted in Fig. X(a)). For auditability, the Gateway/Storage emits a Used event with a commitment to the ticket and object digest.

Optional FHE compute path (permit → compute → commitment). When the grant's opMask includes COMPUTE, the contract can issue a short-lived JobPermit bound to (datasetId, algorithmId, purposeHash, expiry, maxRuns, regionCode). A Compute Orchestrator watches JobPermit events, executes the requested workflow over ciphertexts under the Owner's FHE context, uploads the encrypted result, and records a Post JobResult on chain — i.e., (resultCommitment, resultURI) — which triggers settlement and completes the audit trail (Fig. X(b); the encrypted-storage swimlane is omitted there for visual simplicity). Decryption keys never leave the Owner; if policy authorizes direct delivery to the Grantee, the Owner supplies key-switching material or uses a multi-key FHE workflow; otherwise, the Owner decrypts and shares permitted derivatives. The event-driven orchestrator loop is shown in Listing S2.

Security and compliance considerations. This design ensures (i) least-privilege access (fine-grained scope and temporal bounds), (ii) purpose limitation and residency encoded as verifiable clauses, (iii) revocability prior to use, (iv) full auditability via immutable events, and (v) data minimization (only commitments/pointers on chain). Because FHE computations operate entirely over ciphertexts, compute providers need not be trusted with plaintext. Keys remain under the Owner ’ s control, aligning with regulatory expectations for data sovereignty and consent withdrawal.

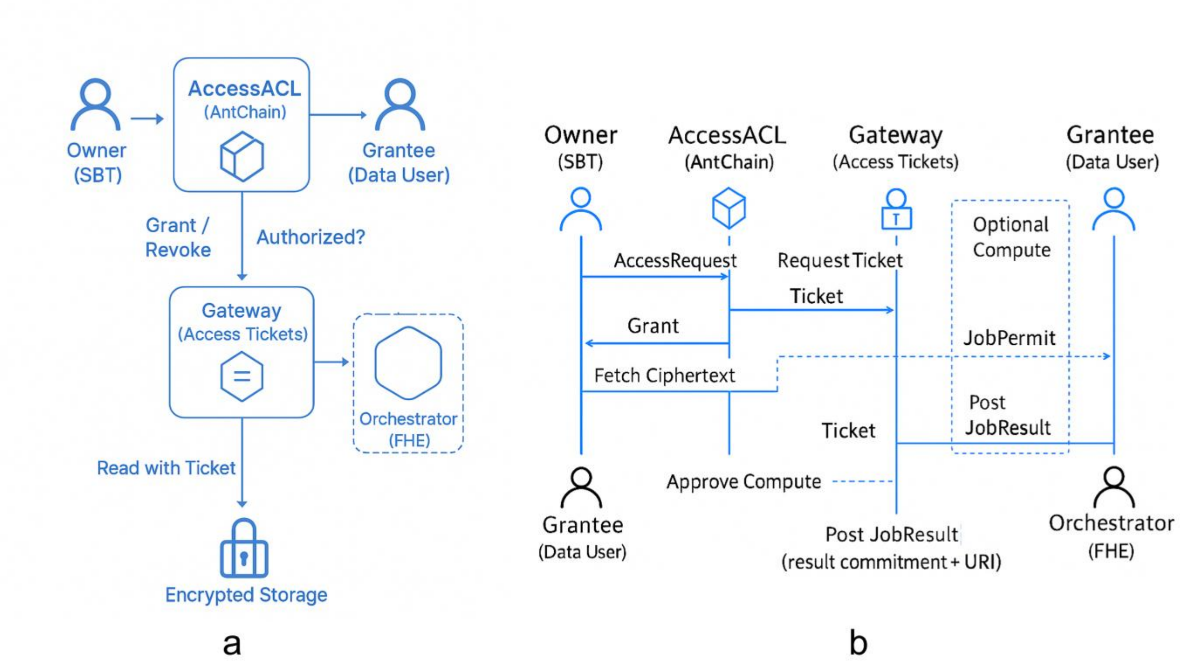


Figure X. Smart-contract access control and FHE workflow on AntChain: (a) system components and data flow; (b) permit→ticket→compute (FHE) sequence.

Data Encryption: Enhancing Data Privacy and Enabling Complete Owner Control

However, blockchain alone is insufficient to protect data due to the inherent nature of digital assets, where duplication costs are virtually zero. Once data are released publicly, it becomes impossible to maintain control over them, even if data ownership is recorded on blockchains. Privacy concerns and restrictive government policies further compound the challenges of sharing data in plain formats.

Various cryptographic strategies have been explored for secure data processing. Here, we will focus on five technologies, including Fully Homomorphic Encryption (FHE), Trusted Execution Environments (TEEs), Federated Learning, Secure Multi-Party Computation (SMPC), and Zero-

Knowledge Proofs (ZKP). Each offers unique advantages and has specific limitations as described in **Supplementary Table 1**.

FHE is a form of encryption that allows computations to be performed directly on encrypted data (ciphertexts) without needing to decrypt it first (**Figure 3**). The result of these computations remains encrypted, and when decrypted by the data owner, matches the result of operations performed on the plaintext[19,20]. This enables secure data processing in untrusted environments such as cloud platforms. TEEs are hardware solutions that provide secure areas within a processor that isolate code execution and data processing from the rest of the system[21,22]. They protect sensitive computations from being accessed or tampered with by unauthorized parties, even if the operating system or other software is compromised. Federated Learning is a machine learning paradigm where a global model is trained collaboratively across multiple decentralized devices or servers holding local data samples, without exchanging the data itself[23,24]. Each participant trains the model on their local data and shares only the model updates (e.g., gradients) with a central server that aggregates them to improve the global model. SMPC is a set of cryptographic protocols that enable multiple parties to jointly compute a function over their inputs while keeping those inputs private[25,26]. Each party learns only the output of the computation and gains no additional information about the other parties' inputs. ZKPs are cryptographic protocols that allow one party (the prover) to prove to another party (the verifier) that a certain statement is true without revealing any additional information beyond the validity of the statement itself[27,28]. ZKPs enable verification of knowledge or attributes without disclosing the underlying data.

Each of these technologies offers a unique approach to addressing data privacy challenges, and their combined use can ensure the highest level of data protection tailored to specific application scenarios. For example, private keys used to decrypt FHE results can be securely stored within a Trusted Execution Environment (TEE), preventing unauthorized access[29]. Among these technologies, Fully Homomorphic Encryption (FHE) stands out as the “holy grail” of encryption, offering a critical advantage: end-to-end data confidentiality, which entirely eliminates exposure of plaintext data. Unlike other methods, FHE does not rely on external parties, hardware manufacturers, or other participants. It allows computations to be securely outsourced to untrusted third-party servers or cloud providers without risking data exposure. By keeping data encrypted during processing, FHE also simplifies compliance with data protection regulations like GDPR, HIPAA and PIPL.

Initially proposed by Craig Gentry in 2009[30], FHE is still in its early stages of implementation. The primary challenge for FHE is its computational overhead, with operations being several orders of magnitude slower than plaintext computations[31]. Addressing this challenge requires an integrative approach that includes advanced scheme and algorithm development, optimized bootstrapping techniques, and hardware acceleration. Selecting the appropriate FHE scheme based on application needs can significantly improve performance[19]. For instance, the Brakerski-Fan-Vercauteren (BFV, 2012) scheme is optimized for integer arithmetic operations. The Torus

FHE (TFHE, 2016) scheme focuses on high-speed bootstrapping and is ideal for binary computations, while the Cheon-Kim-Kim-Song (CKKS, 2017) scheme is designed for approximate computations on floating-point numbers, making it particularly useful for machine learning applications. Hardware acceleration is another critical factor in enhancing FHE efficiency[31]. The development of Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) can greatly speed up FHE operations. ASICs, for example, are estimated to accelerate computations by more than 10,000 times compared to single-core CPU[32]. Additionally, application-specific optimizations, such as pre-computed lookup tables, can reduce computational requirements, and layered security models can be used to categorize data according to privacy levels, optimizing processing based on the sensitivity of the data.

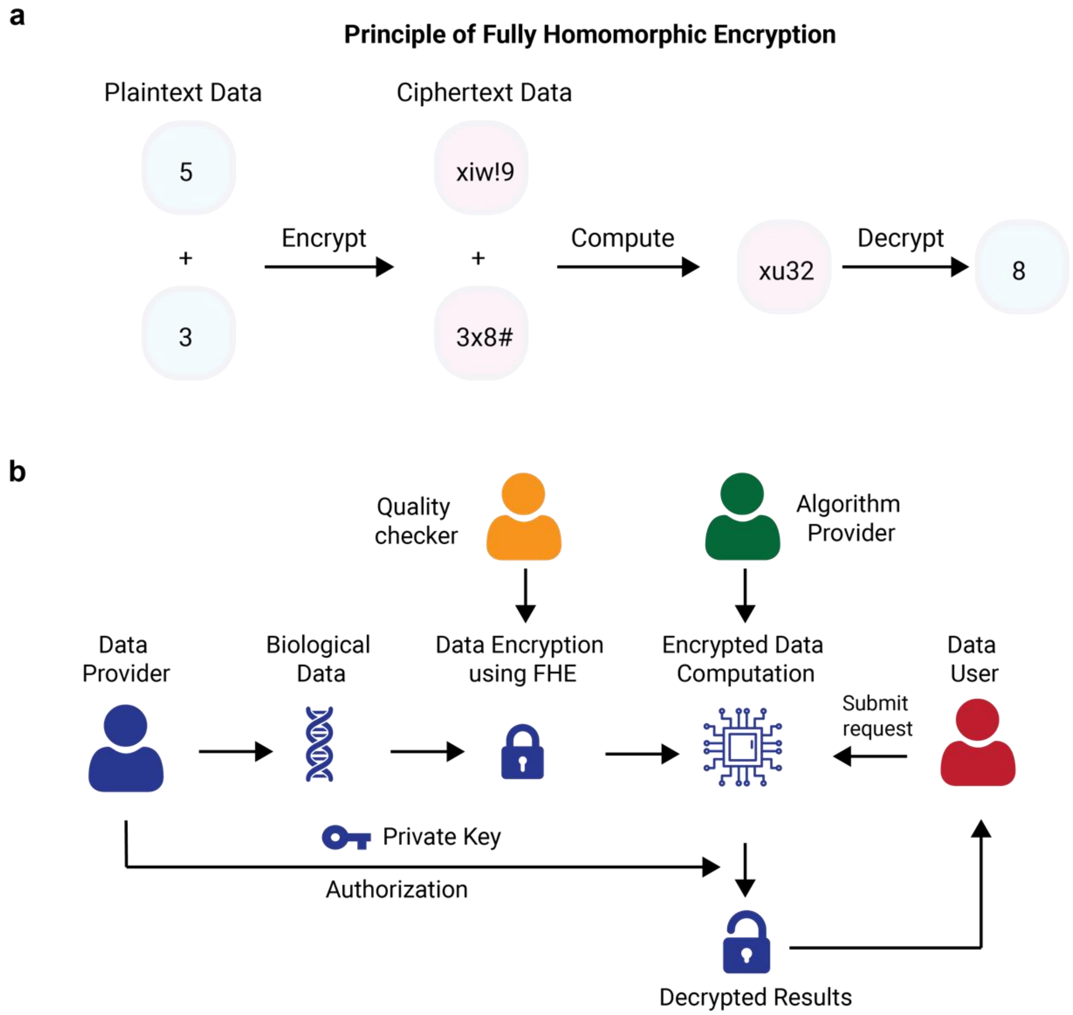
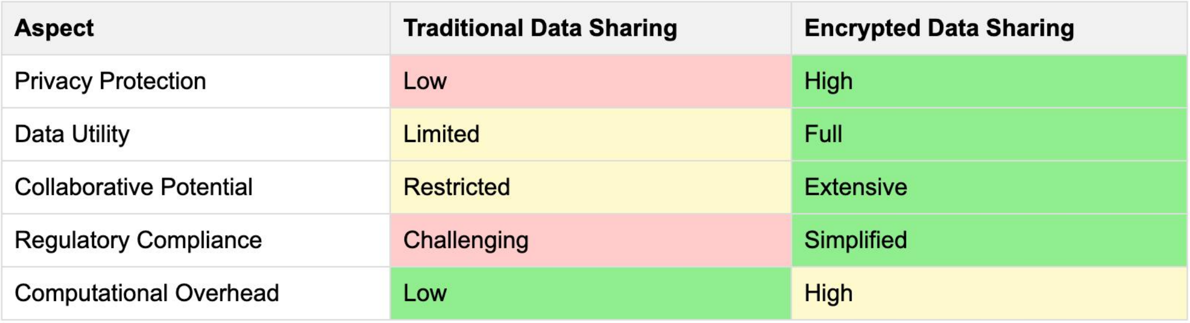


Figure 3. a) The principle of Fully Homomorphic Encryption (FHE) for data encryption and computation. b) Illustration of complete data owner control over data usage using FHE, ensuring end-to-end data confidentiality. The data provider encrypts their data using FHE, and the encrypted data undergoes a quality check. Data users submit requests to perform computations on the encrypted data

using algorithms supplied by an algorithm provider. Importantly, the results of these computations can only be decrypted with the private key held by the data provider.

The sensitivity of biomedical data, particularly genetic data, raises significant concerns about data safety and privacy. Genetic information not only reveals an individual's health status but also discloses family genetic history, potential disease risks, and personal identifiers. Studies have shown that even anonymized genetic data can be re-identified through cross-referencing with other publicly available data sources[33,34]. Recent advancements in Fully Homomorphic Encryption (FHE) have significantly enhanced the potential for secure and privacy-preserving analysis of genomic and clinical data at scale, offering significant advantages over traditional data sharing methods (**Table 1**)[35-39]. For example, Cheon et al. who developed an FHE-based method for genome-wide association studies (GWAS), and their approach allows chi-square statistics to be computed on encrypted genomic data, achieving accuracy comparable to traditional methods while preserving data privacy [36].



Aspect	Traditional Data Sharing	Encrypted Data Sharing
Privacy Protection	Low	High
Data Utility	Limited	Full
Collaborative Potential	Restricted	Extensive
Regulatory Compliance	Challenging	Simplified
Computational Overhead	Low	High

Table 1. Comparison of Traditional and Encrypted Data Sharing Approaches. This table compares traditional and encrypted data sharing approaches across five key aspects. Color coding indicates the relative strengths and weaknesses: red for low/poor, yellow for medium, and green for high/good performance in each aspect. Encrypted data sharing shows clear advantages in privacy protection, data utility, collaborative potential, and regulatory compliance, while traditional methods have an edge in lower computational overhead.

Blockchain- and FHE-integrated Biomedical DePIN devices

To further enhance data security, we propose developing blockchain- and FHE-integrated DePIN (Decentralized Physical Infrastructure) devices such as DNA sequencers that perform FHE encryption and blockchain ownership recording at the point of data generation. This ensures that genetic sequencing data is protected from the moment it is created, greatly reducing the risk of data leakage. This approach not only safeguards user privacy but also builds public confidence in genetic sequencing technologies, encouraging broader participation in genomic research.

Collaborations with leading sequencing device manufacturers such as Illumina, Oxford Nanopore, and BGI could facilitate the integration of such technologies into existing devices.

FHE-compatible analysis algorithm hubs

To ensure the usability of encrypted data, standardized analysis workflows and APIs tailored to common bioinformatics tasks are essential. We propose developing a suite of encrypted data analysis tools, including those for GWAS, epigenomics, transcriptomics, proteomics, and multi-omics integration. These tools will abstract the complexity of encrypted computations, providing researchers with familiar interfaces and output formats. This approach enables biologists who may lack expertise in encryption technologies to conduct sophisticated data analyses effortlessly. We also envision creating an encrypted bioinformatics platform similar to Galaxy[40], equipped with a graphical user interface, workflow management, visualization tools, integrated development environment, data management system, computational resource management, and collaboration tools. This platform will significantly lower the technical barrier to using encrypted data, fostering wider participation in secure data sharing and analysis. Standardized workflows will further reduce these barriers, promoting broader scientific collaboration.

Crypto Token-based incentive system for high-quality data sharing

Addressing data privacy concerns alone does not automatically result in active participation in data sharing. To further incentivize the sharing of high-quality datasets, we propose a crypto token-based incentive system, demonstrated here under the name OMICS (an umbrella term encompassing biological and medical data) as a conceptual framework (**Figure 4**).

In the proposed OMICS system, data providers submit information about their encrypted datasets onto a blockchain platform and mint a Non-Fungible Token (NFT) representing each unique dataset. This NFT serves as a digital certificate of ownership and provenance, ensuring the dataset's integrity and authenticity within the decentralized network. The NFT can then be fractionalized into fungible tokens, enabling divisible ownership or profit rights associated with the dataset. A portion of these fungible tokens is allocated back to the data provider, granting them continued stake and potential revenue from their data asset. The remaining tokens are listed on decentralized markets or exchanges for trading, providing liquidity and allowing investors or other researchers to acquire stakes in the dataset.

Beyond capital raised through token trading, there are three primary mechanisms to generate additional funding and enhance the value of the tokens:

1. *Upfront Submission Rewards from the Platform*: the upfront submission reward is provided to data providers for submitting datasets and passing quality checks. These rewards are issued in OMICS tokens by the platform and tiered based on data quality and completeness. The data quality can be reviewed through a crowd-source peer review system by peer

scientists who have expertise in the corresponding domain and the peer scientists who complete the review can also obtain OMICS token awards.

2. *Pay-Per-Use Payments*: the pay-per-use payment system allows data consumers can access and perform computations on the encrypted datasets by paying usage fees facilitated through smart contracts. These payments are automatically and transparently distributed to token holders, providing an ongoing revenue stream linked directly to the dataset's utilization.
3. *Long-Term Credit-Sharing Royalties*: the credit-sharing royalty mechanism tracks the profits gained by data users from utilizing shared data, such as revenue from publications, patents, or product commercialization. Through smart contracts, a predetermined percentage of these profits is distributed to the original data providers and token holders using OMICS tokens in proportion to their data’s contribution. Data users are required to declare revenues generated from shared datasets through the platform. This information is cross-verified using blockchain-based data usage records and publicly available records, such as patent databases or publication citations. Non-compliance or detected underreporting can result in penalties, suspension, or exclusion from the platform. This ensures that data providers continue to benefit from the long-term value generated by their data.

The abovementioned system is often challenged by limited dataset size, variable dataset quality, lack of funding for data curation, ambiguous data pricing strategy and unclear identification of data buyers. We here propose a collaborative dataset curation mechanism that allows individuals, organizations, or investors to crowdfund or invest in the creation, curation, quality-control and commercialization of datasets in a community-driven, decentralized, and transparent manner.

The workflow is structured into five key phases: *proposal, funding, dataset curation, dataset release, and profit-sharing*. This organized approach promotes transparency, accountability, and fair rewards for all stakeholders. Notably, the incentive system is not a one-directional, fixed process; it is dynamic. As token prices increase, they create a positive feedback loop that drives greater participation and engagement from stakeholders across all stages, reinforcing the ecosystem's growth and sustainability.

1. *Proposal Phase*: A proposer submits a comprehensive dataset proposal, detailing the scope (description, use cases, and scientific objectives, such as rare disease genomics), funding goals (budget requirements), milestones (phased deliverables and timelines), and contributor incentives. Contributors may include data providers, funding supporters, quality checkers, annotators, brokers, community operators, and the platform itself. These elements form the foundation for a community-driven dataset creation process.

2. *Funding Phase*: The project is launched on the platform, enabling investors to pledge OMICS tokens during the offering period. In return, investors receive dataset-specific

tokens, which grant equity or utility rights such as access, profit-sharing, and voting power in the dataset's governance. These tokens can be traded on the market, with a portion of the proceeds allocated to the project's funding pool.

3. *Dataset Curation Phase*: Funds raised are used to recruit data providers, collect data, ensure quality control and proper dataset annotation. Data providers receive tokens as compensation, though the tokens are locked until the data passes quality checks. Funds are released in tranches upon milestone completion, with each milestone verified and recorded on the blockchain. The dataset may establish a Decentralized Autonomous Organization (DAO) to govern the curation process and manage funding allocation, with token holders actively participating in decision-making, such as setting data quality standards.

4. *Dataset Release Phase*: The curated dataset is encrypted and securely stored on the platform. Additional stakeholders, such as data brokers, may be introduced to promote and negotiate usage deals with data consumers, expanding the dataset's reach and facilitating its commercialization.

5. *Profit-Sharing Phase*: Revenues generated from dataset usage (e.g., pay-per-use fees, licensing royalties, or downstream commercialization) are distributed among token holders. A transparent revenue-sharing mechanism allocates earnings proportionally to token ownership. Additionally, a token burning mechanism reduces token supply, potentially increasing token value over time. All transactions and revenue flows are immutably recorded on the blockchain, ensuring transparency, fairness, and trust for all contributors and investors.

The incentive system can also be designed to reduce malicious behaviors. A reputation system can be implemented, where users earn reputation points based on their contributions, behavior, and the quality of their data. Higher reputation levels would unlock additional benefits, such as increased OMICS rewards or reduced transaction fees, incentivizing users to maintain high standards of participation and collaboration. Additionally, a staking mechanism can be introduced, allowing users to stake OMICS tokens on the platform in exchange for rewards that are proportional to the amount staked. This mechanism encourages users to actively participate and align their interests with the platform's success. Any malicious behavior or breach of platform rules would result in the forfeiture of staked OMICS and a corresponding reputation score drop, creating a deterrent against unethical actions and fostering a culture of accountability. Staking OMICS tokens could also grant users voting rights in platform governance, particularly if the platform operates as a Decentralized Autonomous Organization (DAO). This empowers users to participate in decision-making processes, ensuring that the platform evolves in a decentralized and community-driven manner. Moreover, the staking mechanism not only incentivizes long-term engagement but also contributes to token stability by reducing the circulating supply and encouraging committed, responsible participation.

The crypto token-based incentive mechanism provides a decentralized and transparent approach to funding and managing high-value datasets, offering significant benefits for all stakeholders. For data providers, it ensures fair compensation for their contributions through milestone-based funding, token rewards, and profit-sharing from dataset usage. For investors, it creates opportunities for financial returns and strategic access to curated datasets. Data users benefit from access to high-quality, encrypted datasets with clear pricing and flexible payment options. The platform fosters collaboration through decentralized governance, allowing stakeholders to influence decisions and monitor progress transparently via blockchain.

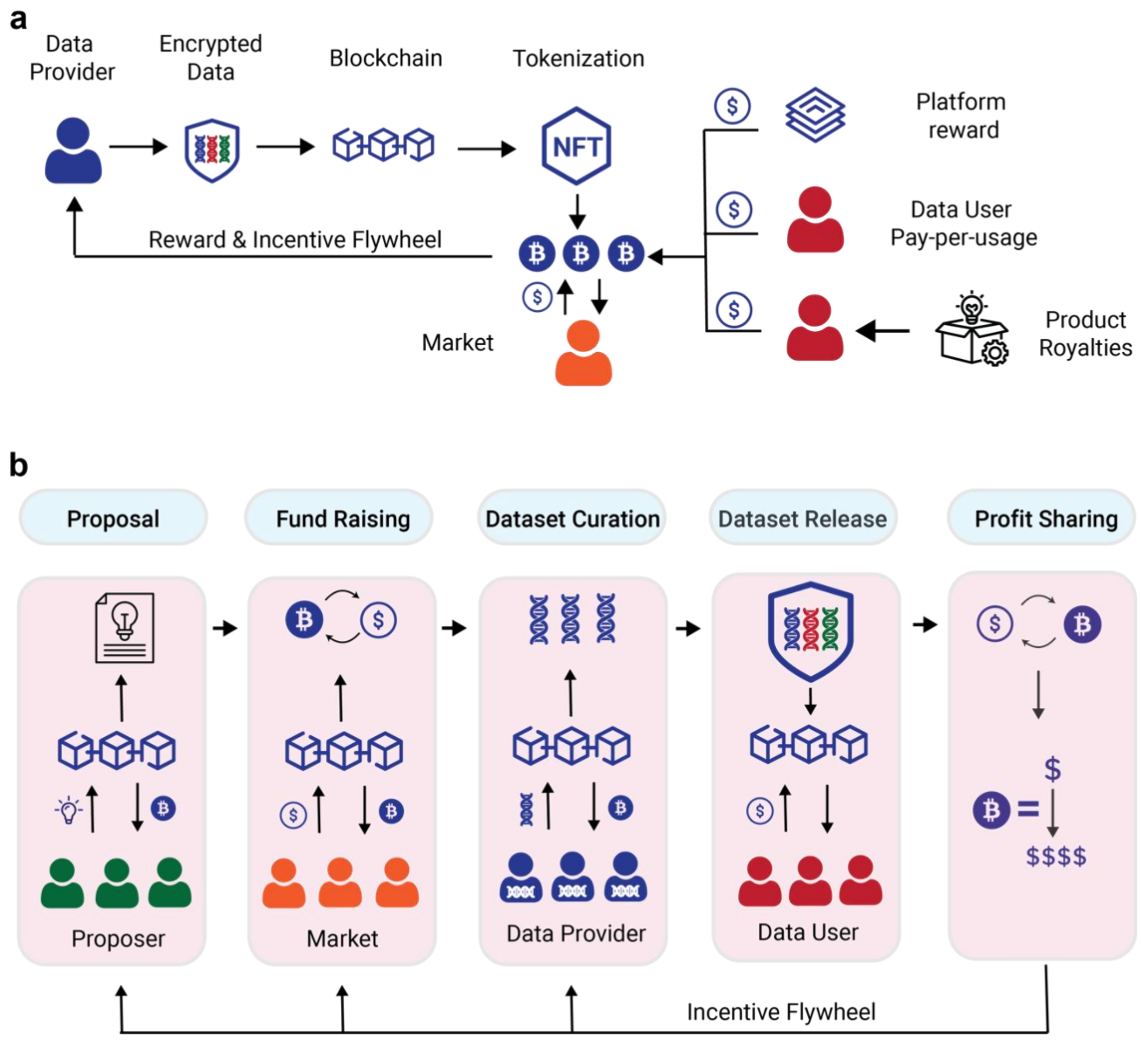


Figure 4. a) A crypto token-based system for dataset tokenization and monetization. The dataset is represented on-chain as a Non-Fungible Token (NFT) and then fractionalized into fungible tokens. A portion of these tokens is allocated to data providers, while the remaining tokens are publicly sold on the market. The value of the tokens is determined by market dynamics and profits generated from platform rewards, data user payments, and royalties. **b) A crypto token-based collaborative dataset curation system.** A dataset concept is proposed, and a corresponding token is minted to represent its value and

potential. Funding for the dataset is raised through the trading of these tokens, which provides the resources necessary for data collection. Data providers are rewarded with tokens in recognition of their contributions. Upon release of the dataset, additional revenues are generated from data users accessing or utilizing the data, further increasing the token's value. This value growth creates a rewarding mechanism for all token holders. The dynamic token market fosters a self-reinforcing flywheel effect, continuously incentivizing the development, fundraising, and data collection process, while aligning the interests of all stakeholders.

It is important to emphasize that the goal of this incentive system is not to create paywalls that restrict access to datasets. Instead, it aims to acknowledge the time, effort, and resources invested by data providers and other contributors, ensuring they receive appropriate recognition and fair compensation for their contributions. Data providers retain the option to donate their datasets freely or waive fees for non-profit researchers. However, this choice should be entirely voluntary, reflecting their autonomy, rather than being imposed by the limitations of traditional systems that fail to protect their rights and interests. This approach strikes a balance between promoting open science and safeguarding the rights of data contributors.

Open Biobank Consortium: a Decentralized Encrypted Biological Data Sharing and Analyzing Network

We propose building a decentralized biological data sharing and analysis network utilizing Web 3.0-related technologies. Achieving secure global biological data sharing also requires a unified approach to establishing technical standards, ethical guidelines, and regulatory frameworks. We suggest forming an international network, tentatively named the "Open Biobank Consortium." This consortium would collaborate with existing organizations such as the Global Alliance for Genomics and Health (GA4GH), Global Biobank Meta-analysis Initiative, UK Biobank, the European Union’s Beyond 1 Million Genomes (B1MG) project, China Kadoorie Biobank, commercial companies such as 23&Me, Wegene and Nebula, and Web 3.0 projects like DataLake and AminoChain. The consortium's goal would be to enhance global management of biological data under FAIR (Findability, Accessibility, Interoperability, and Reusability) principles[41], connecting international efforts to create a unified, secure, and privacy-preserving global data-sharing ecosystem.

Challenges and Prospects

Implementing a Cryptographic Open Science framework offers transformative potential for biomedical research but also presents a range of challenges (**Table 2**). On the positive side, this framework could significantly accelerate the advancement of precision medicine. Secure data sharing will enable researchers to access large-scale, diverse datasets, driving progress in scientific research, drug development, and personalized treatment plans. For example, researchers studying rare diseases often face difficulties in reaching statistically significant conclusions due to limited sample sizes. Encrypted data sharing can overcome geographical and institutional barriers, allowing researchers to pool sufficient case data, thus speeding up the discovery of disease mechanisms and the development of targeted therapies. Additionally, this framework will promote global scientific collaboration, break down data silos, and accelerate the pace of scientific discovery.

Cryptographic Open Science Platform Features	
Feature	**Description**
Zero Trust	Data providers' information is fully protected through Fully Homomorphic Encryption (FHE), allowing them to share data securely without needing to trust other entities.
Per-Use Authorization	Data providers maintain complete control over data usage by authorizing each access with private keys, addressing concerns over loss of control after data is shared.
Provable Privacy	The platform guarantees data privacy through verifiable cryptographic methods, meeting regulatory standards and addressing privacy concerns of oversight bodies.
Transparency and Verifiability	Data ownership, transactions, and usage are recorded immutably on a public blockchain, providing transparency and allowing for verifiable audit trails.
Fair incentives	Market driven incentives to provide funding and compensation to data providers, to enhance data quality and to promote data usage
Automated Credit Sharing	Smart contracts automate transactions, ensuring that credits and benefits are shared equitably with data providers, enhancing incentives for data sharing.

Table 2. Cryptographic Open Science Platform Features.

However, the implementation of this framework also faces several technical, economic, and social challenges.

From a technical perspective, while FHE technology has made significant strides recently, it is still evolving and requires further optimization to handle large-scale biological data analysis effectively. Both academic and industry players are actively developing FHE libraries for practical applications, including IBM’s HELib, Microsoft’s SEAL, PALISADE, and TFHE[42].

Additionally, hardware companies like Intel are working on accelerators to boost FHE performance[31]. Given the current pace of development, we anticipate that FHE performance could reach levels comparable to plaintext computation within the next 5-10 years. Additionally, transitioning existing bioinformatics algorithms to encrypted computing platforms remains a massive engineering challenge, necessitating the redesign of many classic algorithms. This task demands substantial investments in manpower, time, and close collaboration between biologists, computer scientists, and cryptographers. To make Fully Homomorphic Encryption (FHE) more accessible to non-experts, the community can focus on three key areas of development. First, we can create FHE compilers that allow users to write code in high-level languages like Python, which the compiler then automatically maps to encrypted computations. This approach enables developers to leverage FHE without needing specialized cryptographic knowledge. Second, FHE Software Development Kits (SDKs) can be developed to provide user-friendly, low-level APIs that simplify FHE implementation. While this requires developers to have a basic understanding of homomorphic encryption, they are shielded from its most complex aspects. Finally, we can build bioinformatic algorithm libraries specifically designed for FHE, effectively transferring fundamental algorithmic functions into an encrypted environment. Hosting hackathons and programming competitions could further facilitate collaborations between FHE experts and bioinformaticians to develop such libraries.

From economic and social perspectives, establishing and maintaining a global encrypted biological data sharing platform requires significant investment and sustained international cooperation. Balancing the interests of diverse stakeholders and establishing a sustainable operational model are key challenges. Traditional data sharing initiatives like UK Biobank and China Kadoorie Biobank rely heavily on public funding and donations. However, the introduction of decentralized Web 3.0 technologies, such as blockchain and token-based incentives, could revolutionize the funding model by encouraging participation and unlocking market-driven investment from the public. This approach has the potential to raise funds at levels far exceeding traditional methods. Nevertheless, careful consideration of the social impact is crucial. Overcoming public concerns and resistance to genetic data sharing will be essential. While FHE offers robust privacy protection, large-scale public education and outreach will be necessary to build understanding and trust in these technologies. Additionally, an international effort is required to harmonize legal and ethical standards across countries and regions, creating a flexible global regulatory framework that respects local requirements while upholding the core principles of data sharing. An international governing body would be needed to ensure that global projects adhere to these standards.

The Cryptographic Open Science framework provides a promising approach to tackling the privacy and security challenges associated with biomedical big data sharing. By integrating advanced cryptographic technologies, decentralized mechanisms, and international collaboration, this framework can facilitate true "open science" while safeguarding individual privacy and data sovereignty. This approach will not only accelerate the development of precision medicine and public health initiatives but also establish a solid data foundation for the application of artificial intelligence in the biomedical field.

Author contribution

Y.W.: Conceptualization; Investigation; Supervision; Writing - original draft & review & editing; J.F.: Investigation; Writing - original draft & review & editing; Z.B.: Investigation; Writing - original draft & review & editing; L.L.: Investigation; Writing - original draft & review & editing; M.Y.: Funding acquisition; Supervision; Writing – review & editing; G.C.: Supervision; Writing – review & editing.

Funding source

This work was supported by Institute of Digital Medicine at City University of Hong Kong.

Declaration of generative AI in scientific writing

GPT 4.0 was used to improve the readability and language of the manuscript.

Reference

1 Liu, W. *et al.* The 1% gift to humanity: The Human Genome Project II. *Cell Research*, 1-4 (2024).

2 Commission, E. *The EU's open science policy*, <https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en> (

3 NASA. *NASA Open-Source Science Initiative*, <https://nasa-impact.github.io/ossi-website/> (

4 UNESCO. *Open Science: making science more accessible, inclusive and equitable for the benefit of all*, <https://www.unesco.org/en/open-science> (

5 Zastrow, M. Open science takes on the coronavirus pandemic. *Nature* **581**, 109-111 (2020).

6 Scheliga, K. & Friesike, S. Putting open science into practice: A social dilemma? *First Monday* (2014).

7 Micheli, M. Public bodies' access to private sector data: The perspectives of twelve European local administrations. *First Monday* (2022).

8 Ness, R. B. & Committee, J. P. Influence of the HIPAA privacy rule on health research. *Jama* **298**, 2164-2170 (2007).

9 Voigt, P. & Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* **10**, 10-5555 (2017).

10 China, N. P. s. C. o. t. P. s. R. o. *Personal Information Protection Law of the People's Republic of China*, <https://personalinformationprotectionlaw.com/#:~:text=The%20PIPL%20came%20into%20effect,legal%20basis%20and%20disclosure%20requirements.> (2021).

11 IBM. *Cost of a Data Breach Report 2024*, <https://www.ibm.com/reports/data-breach> (2024).

12 Tidy, S. M. J. *23andMe: Profiles of 6.9 million people hacked*, <https://www.bbc.com/news/technology-67624182> (2023).

13 MyHeritage. *Cybersecurity Incident: June 10 Update*, <https://blog.myheritage.com/2018/06/cybersecurity-incident-june-10-update/> (2018).

14 Crosby, M., Pattanayak, P., Verma, S. & Kalyanaraman, V. Blockchain technology: Beyond bitcoin. *Applied innovation* **2**, 71 (2016).

15 Sarmah, S. S. Understanding blockchain technology. *Computer Science and Engineering* **8**, 23-29 (2018).

16 Pinto, R. P., Silva, B. M. & Inacio, P. R. A system for the promotion of traceability and ownership of health data using blockchain. *IEEE Access* **10**, 92760-92773 (2022).

17 Kaaniche, N. & Laurent, M. in *2017 IEEE 16th International Symposium on Network Computing and Applications (NCA).* 1-5 (IEEE).

18 Khan, S. N., Loukil, F., Ghedira-Guegan, C., Benkhelifa, E. & Bani-Hani, A. Blockchain smart contracts: Applications, challenges, and future trends. *Peer-to-peer Networking and Applications* **14**, 2901-2925 (2021).

19 Marcolla, C. *et al.* Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE* **110**, 1572-1609 (2022).

20 Kun, J. *A High-Level Technical Overview of Fully Homomorphic Encryption*, <https://www.jeremykun.com/2024/05/04/fhe-overview/> (2024).

21 Sabt, M., Achemlal, M. & Bouabdallah, A. in *2015 IEEE Trustcom/BigDataSE/Ispa.* 57-64 (IEEE).

22 Jauernig, P., Sadeghi, A.-R. & Stapf, E. Trusted execution environments: properties, applications, and challenges. *IEEE Security & Privacy* **18**, 56-60 (2020).

23 Kairouz, P. *et al.* Advances and open problems in federated learning. *Foundations and trends® in machine learning* **14**, 1-210 (2021).

24 Mammen, P. M. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428* (2021).

25 Zhao, C. *et al.* Secure multi-party computation: theory, practice and applications. *Information Sciences* **476**, 357-372 (2019).

26 Tran, A.-T., Luong, T.-D., Karnjana, J. & Huynh, V.-N. An efficient approach for privacy preserving decentralized deep learning models

based on secure multi-party computation. *Neurocomputing* **422**, 245-262 (2021).

27 Sun, X. *et al.* A survey on zero-knowledge proof in blockchain. *IEEE network* **35**, 198-205 (2021).

28 Berentsen, A., Lenzi, J. & Nyffenegger, R. An introduction to zero-knowledge proofs in blockchains and economics. *Federal Reserve Bank of St. Louis Review* **105**, 280-294 (2023).

29 Behnke, R. *What Is a Trusted Execution Environment (TEE)?*, <https://www.halborn.com/blog/post/what-is-a-trusted-execution-environment-tee> (2023).

30 Gentry, C. in *Proceedings of the forty-first annual ACM symposium on Theory of computing.* 169-178.

31 Zhang, J. *et al.* Sok: Fully homomorphic encryption accelerators. *ACM Computing Surveys* (2022).

32 Ahmad Al Badawi, D. B. C., Yuriy Polyakov, and Kurt Rohloff. (2023).

33 Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nature genetics* **52**, 646-654 (2020).

34 He, Z. & Zhou, J. Inference attacks on genomic data based on probabilistic graphical models. *Big Data Mining and Analytics* **3**, 225-233 (2020).

35 Blatt, M., Gusev, A., Polyakov, Y., Rohloff, K. & Vaikuntanathan, V. Optimized homomorphic encryption solution for secure genome-wide association studies. *BMC Medical Genomics* **13**, 1-13 (2020).

36 Kim, D. *et al.* Privacy-preserving approximate GWAS computation based on homomorphic encryption. *BMC Medical Genomics* **13**, 1-12 (2020).

37 Yang, M. *et al.* TrustGWAS: A full-process workflow for encrypted GWAS using multi-key homomorphic encryption and pseudorandom number perturbation. *Cell Systems* **13**, 752-767. e756 (2022).

38 Grishin, D., Obbad, K. & Church, G. M. Data privacy in the age of personal genomics. *Nature biotechnology* **37**, 1115-1117 (2019).

39 Grishin, D. *et al.* Citizen-centered, auditable and privacy-preserving population genomics. *Nature Computational Science* **1**, 192-198 (2021).

40 Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic acids research* **48**, W395-W402 (2020).

41 Jacobsen, A. *et al.* Vol. 2 10-29 (MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., 2020).

42 Gouert, C., Mouris, D. & Tsoutsos, N. Sok: New insights into fully homomorphic encryption libraries via standardized benchmarks. *Proceedings on privacy enhancing technologies* (2023).

43 Ryan Gifford, J. G., and Joseph Wilson. *Understanding the Differences Between Fully Homomorphic Encryption and Confidential Computing*, <https://cloudsecurityalliance.org/blog/2024/08/22/understanding-the-differences-between-fully-homomorphic-encryption-and-confidential-computing> (2024).

44 Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A. & Qadir, J. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine* **158**, 106848 (2023).

45 Wen, J. *et al.* A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics* **14**, 513-535 (2023).

46 Ray, A. *Challenges Of Zero-Knowledge Proof Technology For Compliance*,

<https://www.forbes.com/councils/forbesbusinesscouncil/2023/08/30/challenges-of-zero-knowledge-proof-technology-for-compliance/> (2023).

Supplementary Tables

Technology	Advantage	Disadvantage
Fully Homomorphic Encryption (FHE)	• Preserve data privacy during computation • Enables secure outsourcing and cloud computation • Support multiple operations on encrypted data	• High computational overhead[43] • Complex implementation[43]
Trusted Execution Environment (TEE)	• Hardware-level security • Higher performance • Compatibility with existing applications	• Requires trust in hardware • Limited resources with constraints on memory and computational capacity • Vulnerability to side-channel attacks that

		exploit physical properties (e.g., timing, power consumption) to extract sensitive data[43]
Secure Multi-Party Computation	• Preserve data privacy during computation • Flexible applications in a wide range of functions and computations	• Computational demanding and require multiple rounds of interaction, leading to high communication overhead[44] • Scalability limitations with performance degrading as the number of participants increases
Federated Learning	• Data privacy compliance with privacy regulation by keeping data within certain geographical regions • Distributed computing to leverage the computational power of edge devices	• Communication overhead consuming significant network bandwidth by frequent transmission of model updates[45] • Security risks with model updates leaking information about local data through reconstruction or inference attacks[45] • Device heterogeneity with variations in device capability and availability affecting training efficiency and model consistency[45]
Zero-Knowledge Proofs	• Data privacy compliance with privacy regulation by keeping data within certain geographical regions • Distributed computing to leverage the computational power of edge devices	• High computational overhead and complex implementation[46] • Provide only proofs and not designed to perform computations on data

Supplementary Table 1. Comparison of different cryptographic technologies for privacy-preserving data computation.

Supplementary Listing S1. Minimal contract–gateway checks.

```
record Grant {
  sbtId; grantee; datasetId; opMask; purposeHash;
  notBefore; expiry; maxRuns; regionCode; termsHash;
}

function canAccess(sbtId, grantee, datasetId, op) view returns bool {
  g = grants[sbtId][grantee][datasetId];
  require(now >= g.notBefore && now <= g.expiry);
  require((g.opMask & op) != 0);
  require(region_ok(g.regionCode));
  return !revoked[g] && within_budget(g);
}

# Gateway side (ticket issuance)
ticket issue_ticket(grantee, sbtId, datasetId, op) {
  assert(canAccess(sbtId, grantee, datasetId, op));
  return sign_gateway({sub:grantee, datasetId, op, exp: now+900});
}
```

Supplementary Listing S2. FHE job permit and orchestrator loop.

```
# On-chain permit creation (owner approves compute)
permit = issue_job_permit(
    sbtId, grantee, datasetId, algorithmId,
    purposeHash, expiry, maxRuns, regionCode, termsHash
)

# Off-chain orchestrator
for ev in watch_chain(["JobPermitIssued"]):
    p = get_permit(ev.permitId)
    if not residency_ok(p.regionCode):
        continue
    ct = fetch_ciphertexts(p.datasetId)            # ciphertexts only
    enc_result = run_fhe(p.algorithmId, ct, owner_ctx(p.sbtId))
    uri, commit = store_encrypted(enc_result)
    tx_job_result(p.permitId, commit, uri)         # triggers audit/settlement
```