Leonardo Toshinobu Kimura

Amazon Biobank: a Blockchain-based Genomic Database for Bioeconomy

Corrected Version

São Paulo, SP

2023

Leonardo Toshinobu Kimura

Amazon Biobank: a Blockchain-based Genomic Database for Bioeconomy

Corrected Version

Master Dissertation presented to the Department of Computer Engineering at Escola Politécnica, Universidade de São Paulo, Brazil to obtain the degree of Master of Science

Universidade de São Paulo – USP

Escola Politécnica

Departamento de Engenharia de Computação e Sistemas Digitais (PCS)

Supervisor: Marcos Antonio Simplício Junior

São Paulo, SP

2023

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 17 de Outubro de 2023

Assinatura do autor: LTKimura

Assinatura do orientador: Markus

Catalogação-na-publicação

Kimura, Leonardo Toshinobu
Amazon Biobank: a blockchain-based genomic database for bioeconomy / L. T. Kimura -- versão corr. -- São Paulo, 2023.
69 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Segurança de Computadores 2.Sistema Distribuído 3.Biotecnologia 4.Blockchain 5.Banco de Dados Genético I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Agradecimentos

Não tenho como agradecer o suficiente meu orientador, Marcos Antonio Simplício Junior, que me apoiou em cada momento de alegria e desafio que passei durante os anos de mestrado. Esteve sempre comigo, seja quando ainda não sabia se ia fazer mestrado, na iniciação científica, nos meus primeiros passos de pesquisa, e em momentos marcantes da minha vida como falecimentos, casamentos, e nascimentos. É um exemplo de profissional, tanto em pesquisa, gestão, ou ensino.

Não consigo agradecer também o apoio da minha família: minha amada mãe, meu falecido pai, meus irmãos, e a minha noiva e atual esposa, Elisângela. Desde antes do mestrado, me acompanharam nessa jornada de diversas maneiras.

Agradeço também aos meus amigos e colegas que me orientaram e aconselharam. Aos meus veteranos, Erina e Valença, pelos exemplos e conselhos. Aos meus colegas da graduação, que mantenho perto até hoje. E aos meus professores e colegas de projeto, entre eles a profª Tereza Carvalho, o prof Wilson Ruggiero, e o prof Ewerton Rodrigues, que tive o privilégio de trabalhar e aprender juntos.

E finalmente, agradeço à Ripple (Ripple's University Blockchain Research Initiative), o CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, processo - 130591/2021-1), e a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil - Código de Financiamento 001), que forneceram o apoio financeiro para realizar esse trabalho.

"A human being in perfection ought ... never to allow passion ... to disturb his tranquillity. I do not think that the pursuit of knowledge is an exception to this rule. If the study to which you apply yourself has a tendency to weaken your affections and to destroy your taste ..., then that study is certainly unlawful..."

Mary Shelley, Frankenstein

Resumo

A bioeconomia é proposta como uma alternativa sustentável para o desenvolvimento econômico de regiões de alta biodiversidade. Especialmente no campo da genômica, o desenvolvimento de sequenciamento genético de baixo custo abre uma oportunidade para novos atores além da academia de se envolverem na atividade. No entanto, os repositórios genômicos existentes sofrem com a falta de rastreabilidade de dados e mecanismos econômicos de compartilhamento de benefícios, resultando em menor motivação para os provedores de dados. Para enfrentar esse desafio, é apresentado Amazon Biobank, um banco de dados genético comunitário. Ao usar as tecnologias blockchain e peer-to-peer (P2P), é permitido o compartilhamento de dados distribuído e transparente, enquanto os smart contracts garantem o compartilhamento justo de benefícios entre todos os participantes do sistema. Além disso, o Amazon Biobank foi projetado para ser auditável por qualquer usuário, reduzindo a necessidade de administradores confiáveis do sistema. Para validar essa abordagem, foi implementado um protótipo usando Hyperledger Fabric e BitTorrent e avaliamos seu desempenho. Os resultados mostram que o protótipo pode suportar pelo menos 400 transações por segundo em uma pequena rede e que isso pode ser melhorado adicionando novos nós ou alocando recursos computacionais adicionais. Assim, é esperado que o Amazon Biobank sirva como uma ferramenta vital para a pesquisa biotecnológica colaborativa, promovendo o desenvolvimento sustentável em regiões de alta biodiversidade.

Palavras-Chave – Biodiversidade, Distribuição de benefícios, Sequenciamento de DNA, Desenvolvimento sustentável da Amazônia, Blockchain, Contrato inteligente, Banco de dados genético

Abstract

Bioeconomy is proposed as a sustainable alternative for the economic development of high-biodiversity regions. Especially in the field of biodiversity genomics, the development of low-cost DNA sequencing opens an opportunity for new actors beyond academia to engage in genomic sequencing. However, existing genomic repositories suffer from a lack of data traceability and economic benefit-sharing mechanisms, resulting in limited motivation for data providers to contribute. To address this challenge, we present Amazon Biobank, a community-driven genetic database. By leveraging blockchain and peer-to-peer (P2P) technologies, we enable distributed and transparent data sharing, while smart contracts ensure fair benefit-sharing among all system participants. In addition, Amazon Biobank is designed to be auditable by any user, reducing the need for trusted system managers. To validate our approach, we implemented a prototype using Hyperledger Fabric and BitTorrent and evaluate its performance. Our results show that the prototype can support at least 400 transactions per second in a small network and that it can be further improved by adding new nodes or allocating additional computational resources. We expect that Amazon Biobank will serve as a vital tool for collaborative biotechnology research, fostering sustainable development in high-biodiversity regions.

Keywords – Biodiversity, Benefit-sharing, DNA sequencing, sustainable development, Blockchain, Smart contract, DNA database

List of Figures

List of Tables

List of abbreviations and acronyms

API	Application Programming Interface
CDB	Convention on Biological Diversity
CDN	Content Delivery Network
DDBJ	DNA Databank of Japan
EBI	European Bioinformatics Institute
HSM	Hardware Security Module
HTS	High Throughput Sequencing
IP	Intellectual Property
IPFS	InterPlanetary File System
NCBI	US National Center for Biotechnology Information
NIG	Research Organization of Information and System National Institute of Genetics
SDK	Software Development Kit
P2P	Peer-to-Peer
POC	Proof of Concept
UDNP	United Nations Development Programme
TK	Traditional Knowledge
TLS	Transport Layer Security

Contents

1 Introduction

Amazon region, with a vast extension of more than 5 million square kilometers, is considered the most species-rich terrestrial ecosystem in the world (HOORN et al., 2010). Being home to 25% of the world's tree species, it has a similar share of other organisms such as microbes, fungi, and animals (BEECH et al., 2017; STEEGE et al., 2013). This incredible biodiversity evolved over more than 50 million years (HOORN et al., 2010) and provides critical ecological services whose value can reach trillions of dollars (STEEGE et al., 2013). Moreover, this region has been inhabited by human population for thousands of years, having a historical impact on its landscape, domesticating species, and holding a large body of knowledge about how to better explore the forest's natural resources (LEVIS et al., 2017; LEVIS et al., 2018).

The Amazonia Third Way initiative (NOBRE; NOBRE, 2019), conducted by the Amazon 4.0 Institute[1], proposes to explore this outstanding biodiversity for sustainable economic development. It tries to be an alternative to what is called the "first way" and the "second way". The first way consists of destructive and predatory activities, such as mining and agriculture, that stimulates deforestation. The second way comprises conservatory and low-profit activities, such as the extraction of standing products (e.g., Brazil Nuts), which keeps the poverty and land conflicts. The biodiversity-based economy, or bioeconomy (STAFFAS; GUSTAVSSON; MCCORMICK, 2013), is an alternative way to keep both biodiversity and economic development (NOBRE; NOBRE, 2019; NOBRE et al., 2016).

Albert originally proposed to be applied to the Amazon Rainforest, this bioeconomy paradigm also can benefit other biodiversity-rich regions. For instance, the Congo Basin Rainforest, the second largest forest in the world, has as its main activity extractive industrial logging (NASI; BILLAND; van Vliet, 2012). Another example is the biodiversity hotspot in Southern Asia, one of the most biodiverse regions of the world (MYERS et al., 2000). Albeit its society already acknowledges the importance of biodiversity (e.g., in traditional medicine and agriculture), it still has many obstacles to modern biodiversity development (RINTELEN; ARIDA; HÄUSER, 2017).

1.1 Motivation

Biotechnology is an invaluable tool to promote sustainable economic development in high-biodiversity regions (NOBRE; NOBRE, 2019). Its research, both in biodiversity and in the traditional knowledge about it, has already promoted innovations in biomimetic

[1] https://amazonia4.org

engineering, synthetic biology, and the development of new materials, chemical compounds, and biofuels (NOBRE; NOBRE, 2019; SILVA; PEREIRA; MARTINS, 2018; RECH, 2011). One example of the biodiversity potential in the Amazon Rainforest region is the açaí fruit (*Euterpe oleracea*). Besides its nutritional qualities and its flavor (BRONDÍZIO, 2008), it has anti-inflammatory and antioxidant agents indicated for the formulation of anti-aging products (PORTINHO; ZIMMERMANN; BRUCK, 2012). Açaí can even be used to mark teeth bacterial plaque to motivate oral hygiene. (DOMINGUES et al., 2012). It is estimated that those and other applications for açaí have added an annual return of more than $1 billion to the regional economy (NOBRE; NOBRE, 2019).

Despite the potential for bioeconomic development, there are still significant challenges to overcome, especially in the context of surveying millions of species across vast forest areas. For instance, in the Amazon Rainforest, which spans over 5 million square kilometers, such activities require enormous effort. Collaboration with local residents, particularly in species identification, can greatly aid in this process. However, their contributions are often inadequately compensated, discouraging their involvement and undermining the local economy. While the Convention on Biological Diversity (CDB) (GLOWKA et al., 1994), a global agreement for biodiversity conservation and sustainable use, emphasizes benefit-sharing, genomic research has historically been conducted under a colonialist model (LI, 2021). To better promote biotechnology, new methods for data sharing are needed. One potential solution is the development of a collaborative and highly scalable genomic database. This database could be populated by residents of relevant areas while preserving data ownership and providing adequate compensation for their contributions.

Currently, many genomic repositories already support biotechnological research, both for academic and industrial purposes. For example, the US National Center for Biotechnology Information (NCBI) maintains a genomic database (SAYERS et al., 2020), as well as the European Bioinformatics Institute (EBI) in Europe (HARRISON et al., 2021), and the DNA Databank of Japan (DDBJ) (FUKUDA et al., 2021). Nevertheless, those repositories make data publicly available without any kind of usage tracing. This model can facilitate data re-usage, but it does not contribute to adequate sharing of economic benefits. For instance, even if a profitable medicine is developed using genomic data from those repositories, usually those profits are not distributed to the correspondent data provider. As a consequence, even people with easier access to genomic data (e.g., residents of high-biodiversity regions) are not encouraged to contribute. This results in less data variety, and less development in the local bioeconomy.

In this context, blockchain and other P2P technologies have been suggested to bring several benefits to genomic repositories. For instance, blockchain can be used as a transparent and verifiable log of transactions involving digital assets (e.g., DNA data). Also, with the addition of a special-purpose currency, blockchain could assist in fair benefit-sharing between all players of the system. Many studies discuss the benefits of blockchain for genomics (OZERCAN et al., 2018; THIEBES et al., 2020; ALGHAZWI et al., 2022; BEYENE et al., 2022). Some of the opportunities listed include data integrity, data ownership (i.e., the owner controls the data utilization), and decentralization (to avoid a single point of failure or to allow distributed data processing). Thus, many proposals incorporate blockchain in their genomic repositories (see Section 6.3). Nevertheless, most of those proposals focus on human genomics instead of biodiversity. Thus, they lack essential functionalities for adequate benefit-sharing, such as transparent correspondence between genomic data and its data provider.

1.2 Goals and System Requirements

The primary goal of the Amazon Biobank is to facilitate collaborative biotechnology research in regions with ecologically rich ecosystems. This research seeks to enhance existing genetic databases, particularly in terms of equitable benefit-sharing resulting from biotechnology. This involves improving traceability by linking each research project to the DNA data used and associating uploaded DNA data with the identity of the uploader. The system must also be auditable, allowing for independent verification of its correct operation without critically trusting administrators.

For this purpose, we consider some functionalities as essential requirements for Amazon Biobank:

- The ability to collect DNA sequences in different forms (raw, assemblies, or annotated) and upload them into the system together with any relevant metadata (common name, scientific name, where it was collected, information about its common usages, etc.)
- Association of the uploaded DNA sequences with the identity of the uploader, to preserve the latter's rights and to ensure fair compensation. This procedure must be performed in a verifiable manner, i.e., it must not depend critically on the trust deposited in the system entities.
- Provision of capabilities for validating the correctness of inserted data (e.g., that processed DNA sequence corresponds to some previously registered raw DNA data), or at least giving confidence of its correctness (e.g., by means of a reputation system). If misbehavior is detected, suitable penalties should be applied, including the possibility of evicting users from the system.
- The ability to search for specific data among the entries inserted into the system. Searches may be performed either on keywords available as metadata or based on similarity with sequences of interest.

- The possibility of purchasing access rights to data of interest and then downloading it. All actors that helped in making that data available (e.g., by collecting, processing, validating, and/or distributing it) should then be properly remunerated.

A few non-functional requirements are similarly relevant:

- Traceability: the system must provide some level of traceability for biotechnology developments resulting from DNA data stored in the Biobank (e.g., scientific discoveries or intellectual property). This feature promotes the reproducibility of results, which can be reliably traced to Biobank entries. This feature is useful both for academic purposes and to support claims about the prior existence of some data in the Biobank when handling disputes involving data misuse.
- Scalability: the system must be able to handle many users uploading and accessing data stored by the Biobank One challenge for this is that DNA files are usually large (e.g., many gigabytes) and operations on them (e.g., sequencing raw data, or searching for specific sequences) can be very time-consuming. We do not require that a large number of entities participate in the Federation though, as only some authorized entities would maintain the system.

1.3 Contributions

This work describes Amazon Biobank, a community-based genetic database fostering the construction of biotechnology-based assets. By combining blockchain and smart contract technologies, the proposal allows adequate benefit-sharing among participants who collect, insert, process, store, and validate genomic data. It also provides traceability and auditability features, so biotechnology products and research can be easily and transparently traced back to data in the repository. These features are useful, for example, for providing certification of origin, and reproducibility, or when solving disputes involving data usage rights. Finally, by leveraging peer-to-peer (P2P) technologies, the Amazon Biobank creates a highly scalable collaborative computing environment where users can contribute with (and get remunerated for) genomic data and computational, storage, and bandwidth capabilities. This architecture follows a zero-trust approach (ROSE et al., 2020), where the underlying security properties have no critical dependency on the honesty of the system or its users.

Besides its technical interest, the Amazon Biobank is expected to also deliver social impacts when integrated with the developing Amazon Creative Labs (NOBRE; NOBRE, 2019). Managed by the Amazon 4.0 Institute[2], these laboratories seek to prepare

[2] https://amazonia4.org/

and train local communities into exploring high-value bioeconomy opportunities. One of those opportunities consists in acting as a data Collector in the Amazon Biobank initiative and receiving “biocoins” for this task. Another is to promote the creation of local biotechnology-related businesses. This can be achieved by leveraging distributed computing from third parties to analyze genomic data and combining the results with local knowledge for building different applications

Although this work is focused on the Amazon Rainforest region, it can also be applied to other high-biodiversity regions. As long as local-specific regulations are followed, Amazon Biobank has the potential to facilitate the fair and transparent benefit-sharing of bioeconomic research and to promote the sustainable development of those areas. Some adaptations might be necessary though, in order to be adequate to specific local regulations.

1.4 Team

Amazon Biobank is part of the Amazonia 4.0 project, conceived by the Amazonia 4.0 institute. Thus, the system’s design received valuable collaborations from a multidisciplinary team, including biologists, economists, and lawyers. Their input was crucial to assessing the system’s viability, and it significantly influenced Amazon Biobank’s design.

The author was mainly responsible for the technical aspects of Amazon Biobank. Thus, he specified most of the design and architecture of the system and implemented a major part of the prototype and experiments. Nevertheless, the development and the prototype evaluation were done in collaboration with our lab members. Also, the P2P client application used in Amazon Biobank was largely done by (SHIRAISHI et al., 2021).

1.5 Outline

The remainder of this document is organized as follows. Chapter 2 introduces blockchain and P2P concepts used in the entire document. Chapter 3 details the system requirements, system players, and the design decisions taken for fulfilling the requirements. Chapter 4 describes the design and the main component of the proposed architecture. Chapter 5 details the prototype implemented and the experiments conducted to evaluate it. Chapter 6 presents some final considerations, including limitations inherent in our system. Finally, we conclude in Chapter 7 with the main results and next steps. In Appendix A, we included documentation of the prototype’s main operation.

2 Background

This section describes the main concepts of blockchain, including Hyperledger Fabric, and P2P file sharing. Those concepts are the main building blocks of Amazon Biobank.

2.1 Blockchain

Blockchain is defined as a block-based data structure, in which each block is linked to the previous one via hash functions, forming a linear chain of blocks (MIT, 2018). New blocks are built by hashing the block data with the latest block of the chain. Usually, blockchain is replicated between several entities, who jointly ensure data integrity. This results in an append-only data structure that is difficult to tamper and forge.

One of the first applications of blockchain is Bitcoin (NAKAMOTO, 2008), a worldwide digital currency that is not subject to any centralized authority. Bitcoin uses blockchain as a transparent and append-only ledger with all monetary transactions. Those transactions are validated by peers to ensure correctness (e.g., by verifying that the sender has sufficient funds) and prevent double-spending. To order those transactions and sync them between all peers, Bitcoin runs an underlying consensus mechanism (proof-of-work), in which miners are rewarded with some coins. As a result, Bitcoin prevents attempts to manipulate the system, either by sending an invalid transaction (detected by peers), or by modifying a past transaction (evidenced by the blockchain structure).

With the introduction of Ethereum, blockchain started to be used not only to exchange digital assets but also to distribute computation between several nodes. To do so, Ethereum employs smart contracts, code that is deployed and executed on distributed peers and that any user can read, call, or verify. The correctness of this code execution is ensured by consensus mechanisms (BUTERIN et al., 2014). This possibility of running code in a transparent and decentralized way resulted in several financial and non-financial solutions, such as De-fi, tokenization, and gaming (SWAN, 2015).

While Ethereum and Bitcoin adopt a highly open and decentralized model, some blockchains restrict who can read from the blockchain, write into it, or has permission to participate in its consensus protocol (SHRIVAS, 2019). Those blockchains, called permissioned blockchains, commonly require identifying all users interacting with the system, be they system nodes or end-users. Thus, they can easier comply with legal requirements, such as identifying the source and the destiny of all transactions to avoid money laundering. In addition, by specifying that only a small number of well-defined

Table 1 – Amazon Biobank blockchain permissions

Operation	Allowed users
Write	Identified users
Read/Audit	Any users
Consensus Participation	Only system managers

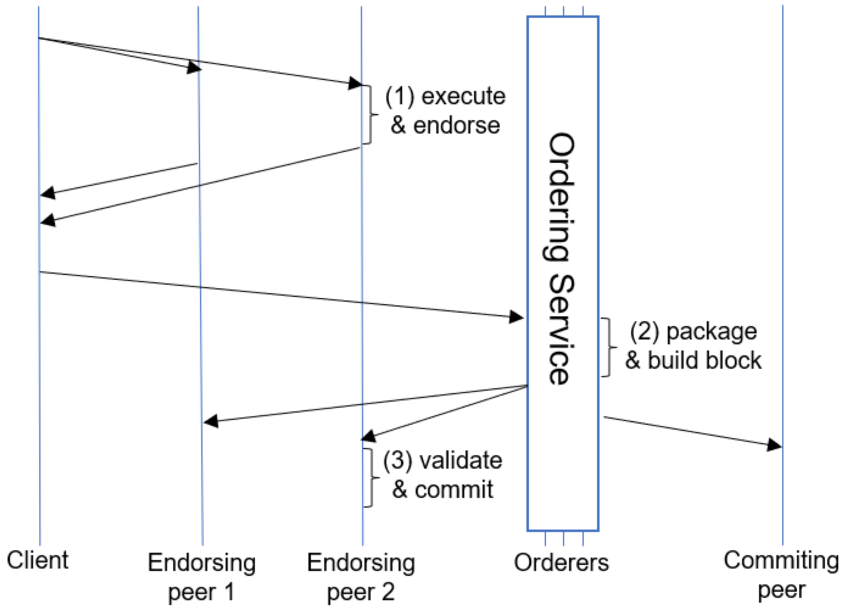
Source: Author

entities can participate in the consensus protocol, permissioned blockchain supports more lightweight consensus mechanisms, like RAFT (ONGARO; OUSTERHOUT, 2014a) or Consensus (CHASE; MACBROUGH, 2018). Those properties make permissioned blockchain attractive for many business applications in an Enterprise context, such as supply chain, medical services, and genomic databases (LI; WONG; GUO, 2020; OZERCAN et al., 2018).

In the context of the Amazon Biobank, the use of blockchain is motivated by the need of providing a transparent and verifiable log of transactions involving digital assets (DNA data and associated metadata) and a special purpose currency (biocoins) (LAURIE, 2014). As further discussed in Section 3.2, the blockchain enables users to not have their data associated with another user after the insertion, even if the latter colludes with malicious system entities. Analogously, with the blockchain, users can verify that the total amount of biocoins involved in the payment of some service is shared appropriately among the players - or, at least, users can easily detect attempts to do otherwise, even in the presence of dishonest system actors. This zero-trust property would be difficult (if possible) to achieve with a regular database controlled by a central entity, which may be compromised or go rogue. Conversely, this property is quite naturally obtained with a (distributed) timestamp authority for registering transactions, which is the role played by the blockchain in the proposed architecture.

In this scenario, a few authoritative entities, like universities and (non-) governmental organizations, act as system managers. Hence, is somewhat natural to use a permissioned blockchain with a lightweight consensus mechanism. Furthermore, since the database does not store any sensitive data (only magnet links for encrypted data, as further discussed in Section 4.1.2), we impose no restrictions on who can read from the blockchain. Therefore, any interested entity can become an auditor, even unregistered ones. All that is needed is the desire to act as a third-party verifier, validating the consistency and the correct operation of the blockchain at any time. However, the system is expected to identify all participating users due to legal requirements, while also thwarting misbehavior (e.g., users uploading fake data into the database). Therefore, only registered users are allowed to send write requests into the blockchain and to vote for proposed system changes whenever required (Table 1).

Figure 1 – High-level transaction flow of Hyperledger Fabric



Source: Author

2.2 Hyperledger Fabric

Hyperledger Fabric is a permissioned blockchain platform, hosted by the Linux Foundation, mainly used in Enterprise applications (LI; WONG; GUO, 2020). It supports flexible smart contracts, also called chaincodes, that are well-suited to a range of applications, inclusive healthcare (ANTWI et al., 2021) or genomics (CARLINI et al., 2019).

When deployed, the Hyperledger Fabric operation is maintained by the Federation, consisting of multiple organizations that independently coordinate to maintain the blockchain operation. To do so, each organization maintains one or more nodes, which can be of the following types: 1) Endorsing nodes, which store and execute the chaincodes and maintains a copy of the blockchain ledger; 2) Ordering nodes, which define the next set of transactions to be included on the blockchain; and 3) the Committing nodes, that does not execute the chaincodes, but store a copy of blockchain ledger and perform signature validations.

This node division between Endorsing nodes, Ordering nodes, and Commiting nodes, is a consequence of the execute-order-validate model of Hyperledger Fabric. While Bitcoin and Ethereum adopt an order-execute model, the Hyperledger Fabric divides the transaction flow into the following steps (see Figure 1):

- *Endorsing phase:* The requesting user sends the transaction to some Endorsing nodes, which make validity checks (e.g., well-formedness, authenticity, and authorizations).

If no errors are detected, those nodes execute the transaction and generate a result, which includes a read-write set that must be stored in the blockchain ledger. Those nodes then sign the result and send it back to the client as an endorsement. The client waits until enough endorsements are collected according to a pre-defined endorsement policy.

- *Ordering phase:* After collecting the endorsements, the client assembles them into a transaction and sends it to the ordering service. This service comprises multiple Orderer nodes, which use a consensus mechanism to coordinate their task. The ordering service packs those transactions into a block and then propagates it to other peers on the network.

- *Commit phase:* After receiving the block from the ordering service, Committing peers make the necessary verification (e.g., transaction format and the endorsement policy). If correct, those peers commit the update on the blockchain according to the transactions. This task must also be done by all Endorsing peers on the network.

This execute-order-validate model allows for flexibility on the employed consensus mechanisms (RAFT, KAFKA), and the use of general programming languages (GO, JAVA, Nodejs) for smart contracts (ANDROULAKI et al., 2018). In addition, it allows organizations to maintain only some types of nodes according to their interest. For instance, organizations that desire to fully participate in the validation process may deploy both Endorsing node and Ordering nodes, while organizations that only want to ensure data immutability may support a more lightweight Committing node. Finally, the endorsing policy, or the number of Endorsing nodes needed to execute the transaction, can be adapted according to the desired application. Some common endorsing policies are the AND policy (requires endorsement from all peers), the OR policy (requires endorsement from at least one peer), and the MAJORITY policy (requires endorsement from most of the peers).

2.2.1 RAFT consensus mechanism

Amazon Biobank uses RAFT as a consensus mechanism, a lightweight, deterministic, and scalable protocol compared to other mechanisms such as proof of work (ONGARO; OUSTERHOUT, 2014b). RAFT uses a "leader and follower" model, in which a leader node is elected to decide the next set of transactions and the follower nodes are responsible for replicating this decision. The follower nodes must also monitor the heartbeat signal of the leader node, executing a new election whenever they detect any unavailability. Therefore, RAFT has Crash Fault Tolerance, being able to support up to 50% crashes on the network. It does not have Byzantine Fault tolerance though, which means that a malicious leader could commit manipulated transactions on the blockchain. Nevertheless, this is not a problem in Amazon Biobank, as permissioned blockchain can simply identify and evict any malicious nodes from the system.

RAFT consensus is projected for small networks; thus, an increase in the number of nodes causes a decrease in its throughput and availability (FU; WEI; TONG, 2021). This happens because larger networks have a higher rate of package loss, requiring longer election phases (HUANG; MA; ZHANG, 2020). However, RAFT can support at least ten nodes without significant loss in performance. As we expect Amazon Biobank to have only a few orderer nodes, the scalability of RAFT consensus is adequate enough.

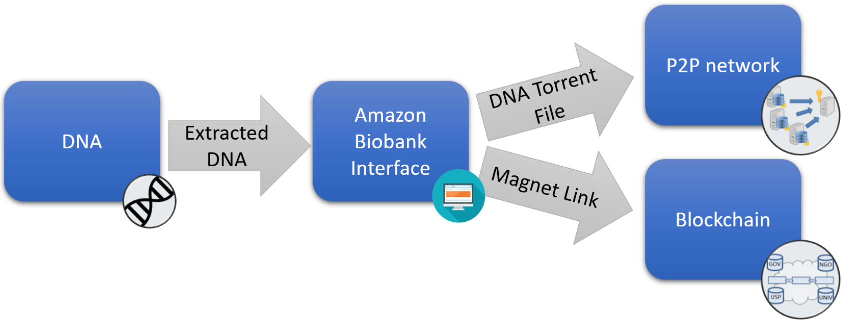
2.3 Bittorrent

BitTorrent (COHEN, 2003) is a communication protocol for peer-to-peer file sharing. Usually used for data transference without the need for a centralized server, BitTorrent is also used by companies to reduce bandwidth requirements of data servers (e.g., Windows Delivery Optimization for large updates (MICROSOFT, 2022)). To download a given data, users must first obtain a .torrent file, containing metadata about the desired data (e.g., file name, size, hashes). This .torrent file can be obtained either from a torrent distribution website or from other peers through a magnet link unique to each file (HAZEL; ARVID, 2008). Then, users must find other peers that contain the desired data. One approach is by contacting a tracker, a centralized entity that keeps track of all peers that have downloaded a specific .torrent file. Another approach is by using the magnet link: users can find each other through a Distributed Hash Table (DHT) (LOEWENSTERN; ARVID, 2008) based on Kademlia (MAYMOUNKOV; MAZIÈRES, 2002). After finding those peers, users must connect to them and download the file, piece by piece. Every piece has its hash stored in the torrent metafile, so its integrity can be verified individually.

While BitTorrent relies on a tic-for-tac incentive mechanism, in which peers contributing to file upload are prioritized (COHEN, 2003), the Torrente protocol proposes an alternative, micropayment-based incentive mechanism (SHIRAISHI et al., 2021). Essentially, in Torrente, each data transference is communicated to a blockchain, and each piece of data is remunerated with an off-chain micropayment. This protocol is somewhat an optimized version of BitTorrent Speed (BitTorrent Foundation, 2019) and is based on a hash chaining mechanism similar to PayWord (RIVEST; SHAMIR, 1997). Thus, peers are remunerated proportionally to the amount of data they have uploaded, encouraging data sharing. This mechanism contrast with other distributed data-sharing methods, like IPFS (BENET, 2014). While in IPFS, peers are remunerated essentially for storing data, in Torrente, peers must primarily upload the data to be remunerated.

In the Amazon Biobank scenario, the adoption of BitTorrent for data sharing is important for creating a more scalable and less centralized solution. In this system, even users with commodity computers can collaborate with the data storage and distribution, while the system manager only needs to keep a copy of the DNA data for ensuring their long-term availability. Therefore, the resulting storage service can be more affordable whilst still highly available and redundant with support from the system users. In addition, Amazon Biobank only stores on the blockchain the magnet links corresponding to DNA data (Figure 2). This results in smaller block sizes, despite the frequently large sizes of DNA files. To avoid the unauthorized distribution of DNA data by peers, all content is stored in encrypted form. Consequently, only users who obtain the decryption key (e.g., by purchasing it) can access the corresponding plaintext.

Figure 2 – High-level steps to upload DNA data in Amazon Biobank



,

Source: Author

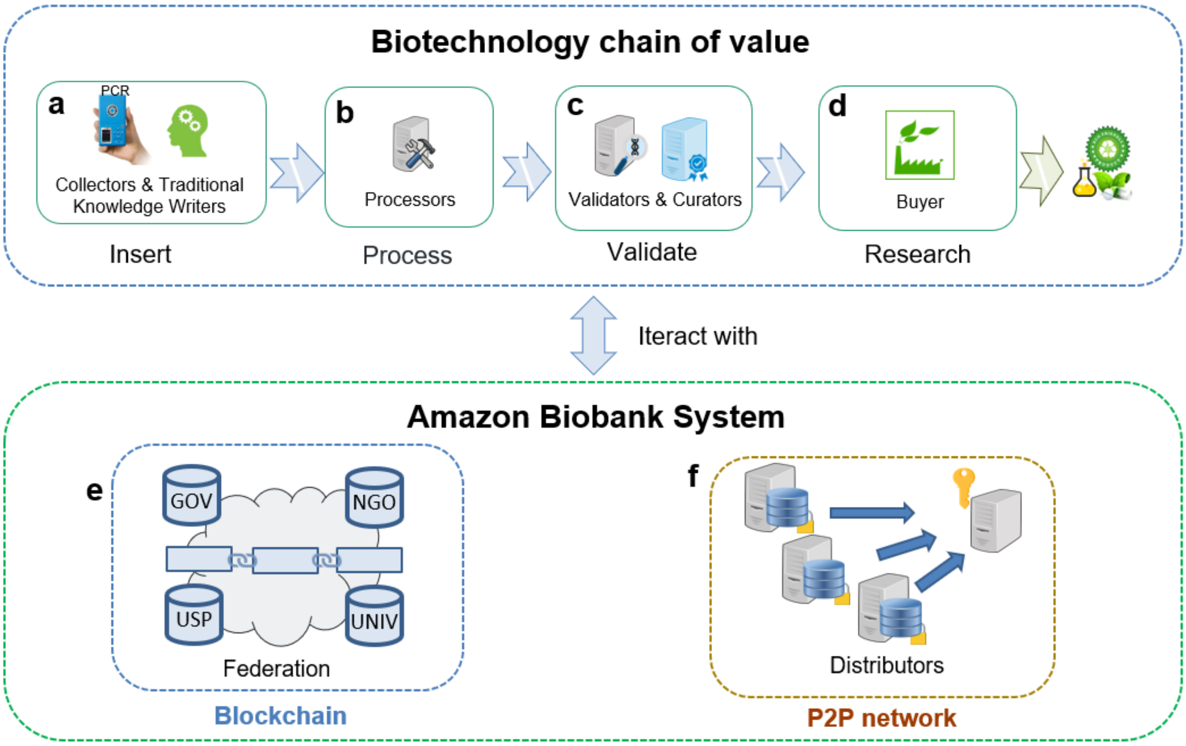
3 The Biobank

In this chapter, we discuss the overall design of Amazon Biobank. We start by describing the main roles played by the system's entities; then, we describe in detail how those players interact to enable the system's operation.

3.1 System Players

The Amazon Biobank system involves the following players, all of which are expected to be authenticated and authorized before they can interact with the system (i.e., we consider a federated environment). We note that those players may represent either natural people or groups of individuals. Figure 3 illustrates those players in Amazon Biobank, and how they interact to enable the system's operation.

Figure 3 – Amazon Biobank: overview and main operations. **a)** Collectors provide raw DNA data, possibly complemented with traditional knowledge files. **b)** Processors sequence and assemble raw DNA. **c)** Validators and Curators assess data correctness. **d)** Buyers purchase access to DNA data, using it for their research. **e)** The Federation comprises universities and (non-)governmental organizations that manage the system. **f)** Distributors store and share data over a BitTorrent-based P2P network



Source: Author

1. Collector: Responsible for collecting raw genomic data in the field, and providing it to the system in the form of a torrent file. This role will usually be played by residents of the target region where the Biobank is deployed, potentially with the support of a proxy (e.g., a university or local Non-Governmental Agency – NGO) for encrypting and delivering the collected data.

2. Traditional Knowledge Writer: A participant responsible for inserting potentially relevant information about the specimen whose DNA data has been collected. This is expected to make the corresponding entries easier to find via keyword search, for example. Data Collectors are likely to play this role themselves, for their own collected data, although other participants may also be authorized to do so.

3. Distributor: A participant whose role is to distribute the data referenced by a torrent file. Every participant in the P2P network who has downloaded the corresponding (encrypted) data can assume this role, contributing to the system's storage and bandwidth capabilities.

4. Buyer: Normally, the Buyer is the entity interested in obtaining access to some piece of DNA data. They can seek specific DNA sequences or associated protein sequences, and download the data of interest, paying the applicable fees.

5. Processor: The system follows a crowd-computing approach, relying on one or more data Processors to offer their own computational power to sequence and assemble raw DNA reads, in exchange for a reward. More precisely, when a Collector uploads raw data to the system, competition among different Processors can occur for processing this data. Once the processing is finished, the Processor registers in the system the torrent file associated with the obtained results, in the form of a magnet link, and gains ownership of this processed data. If two or more Processors submit their results at similar times, the decision on who will be the owner is taken according to some well-defined policy (usually, data is registered in the order of arrival). Processors who invested computational power in this task, but were not chosen as the owner of the processed data, can leverage their efforts as data Validators.

6. Validator: users registered as Processors can also validate the quality of processed data uploaded by other peers, registering positive or negative votes on existing entries. This Validator role is useful for promoting data quality and avoiding attempts to register bogus data.

7. Curator: Formed by NGOs, university members, and biologists working to maintain the system, its role is to ensure that the data inserted in the system and its correspondent metadata are valid. They are also responsible for moderating and resolving conflicts, registering their verdicts into the blockchain whenever necessary.

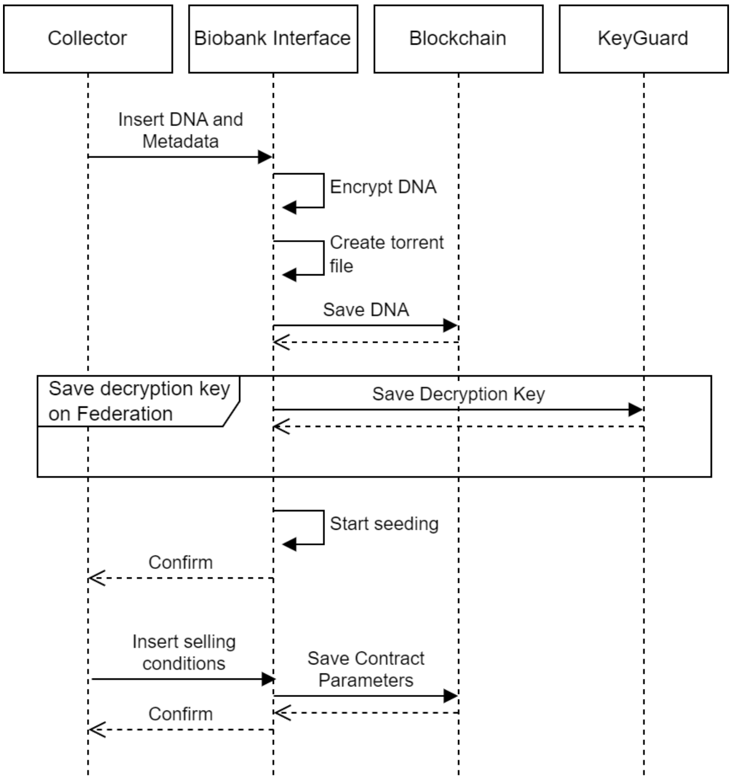
For example, Curators are expected to take action when a given data entry receives multiple positive and negative votes by Validators, or when someone claims that a raw DNA entry is bogus.

8. Federation: the group responsible for managing all system operations, deploying smart contracts, and verifying the credentials of participants whenever necessary. It consists of a small group of entities, including Universities, and (non-)governmental organizations, running a suitable consensus protocol (BACH; MIHALJEVIC; ZAGAR, 2018).

3.2 Register DNA Sequences

One of the Biobank system's primary operations is the registration of raw DNA data (see Figure 4). Essentially, Collectors extract raw DNA data from local species, upload this data to the system, submit some details (e.g., common name, place of extraction) and define payment parameters (e.g., distribution of royalties or profit among system players). As a reward, Collectors receive biocoins when Buyers purchase access to the data when the data itself leads to some profitable product, and when the DNA information is deemed "unique" (i.e., that adds to the overall diversity of the Biobank). This section describes this process in more detail.

Figure 4 – Sequence diagram - Upload DNA data



Source: Author

3.2.1 Inserting genetic data

First, a Collector (e.g., a resident of the Amazon region) collects some DNA sequences using a portable sequencing device, resulting in a raw instrument signal or a DNA sequence read. Then, the Collector encrypts the collected data and creates a torrent file. This torrent file is then uploaded to the system through its magnet link, creating an association between the Collector and the inserted data. As this association is registered in the blockchain, Collectors can be confident that it cannot be "hijacked" by some other user (even in collusion with system administrators).

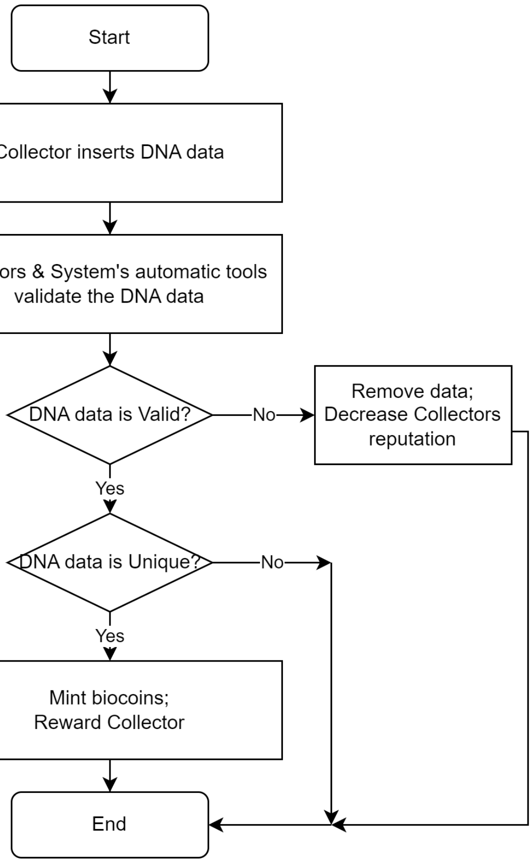
Collectors are expected to also attach useful information to the DNA sequence. This metadata might include, for example, the data origin (e.g., personally collected or taken from other databases), how it was collected, and when, and where. The metadata could also contain the species referred to in the inserted DNA, as well as other information of potential interest, like pictures, popular names, the geographical location where the specimen was collected, and common usages (e.g., due to its medicinal properties). This metadata may be stored without encryption, so Buyers can use this information when searching for data of interest. Alternatively, only general terms may be made public, while detailed information would remain concealed. For example, one might disclose the fact that some plant is associated with “medicinal properties”, while the specific illnesses treated with them are only accessible with the corresponding decryption key.

Finally, Collectors can share the encrypted data with Distributors, who can help with the storage and delivery of data pieces to interested parties. By default, Collectors are expected to share the decryption key with the Federation, thus facilitating the implementation of services like sequence search, as well as data availability control. Nevertheless, Collectors that have enough computational capabilities may prefer to keep the decryption key to themselves, handing search and/or backup services on their own.

3.2.2 Verifying data correctness

Curators may try to, intentionally or maliciously, insert incorrect data or mismatch metadata. Thus, Biobank combines two mechanisms to promote confidence in the registered data. The first is Curatorship, in which players registered as Curators (typically biologists) can verify the correctness of the registered metadata, besides inserting their own (See Figure 5). Curators follow verification guidelines, published and periodically reviewed by the Federation, and provide stamps of approval upon verified DNA data. The second mechanism is a reputation system, in which players who provide correct metadata are rewarded while incorrect metadata are penalized. Users with low reputations may lose access to part or all of their funds, compensating peers affected by their misbehavior and rewarding those who detected the misdeed. The quality of the registered information may also be assessed indirectly, via the reputation of the corresponding Collector. Those two mechanisms, together with some automatic validation mechanisms for assessing genome assembly and annotation completeness (SEPPEY; MANNI; ZDOBNOV, 2019), should at least dissuade malicious activities.

Figure 5 – Flow Chart - Validate DNA data


Start
Collector inserts DNA data
Curators & System's automatic tools validate the DNA data
DNA data is Valid?
No
Remove data; Decrease Collectors reputation
Yes
DNA data is Unique?
No
Yes
Mint biocoins; Reward Collector
End

Source: Author

In addition, to encourage data diversity, the Biobank can reward with fresh-minted biocoins the Collectors who insert unique DNA samples. This mechanism favors Collectors who provide more value-adding information, and it limits the incentives for users who insert redundant DNA data. The responsibility for evaluating uniqueness may involve both: (1) automated mechanisms enabling similarity comparisons with entries in the Biobank and in public databases (GOLDSTEIN; DESALLE, 2011); and (2) manual intervention by Curators. Malicious attempts to pollute the system with duplicated data can be penalized (e.g., by decreasing reputation). We note that this mechanism does not prevent Collectors who insert data for the same species from getting paid for their efforts. After all, Buyers may choose to purchase access rights to the first entry added into the system, to entries that came subsequently, or even to both (e.g., aiming to assess DNA variability in a region).

3.2.3 Defining payment distributions

Collectors must also define how their data can be used by interested parties. For example, they should specify which rules (e.g., royalties) apply, how other players participating in the supply chain (e.g., Processors) are rewarded when access rights to the content are acquired, and any other applicable minutiae. Smart contracts are then registered in the blockchain to enforce all the applicable rules. We note that the system may recommend some default rules, as well as impose some upper and lower bounds to given parameters. For example, when allowing widespread access to their data, Collectors might define that the resulting benefits are shared as follows: 1% to the Federation, for maintaining the system; 30% to the Processor; 30% to a maximum of 5 Validators, and the rest to the Collector him/herself. Nevertheless, Collectors are free to define their own rules within the system's limits. For instance, Collectors could increase the reward for Processors or Validators, aiming to encourage their collaboration or reduce their rewards if the genetic data is believed to be worth processing even without extra incentives (e.g., due to its apparently high commercial value). Naturally, not all inserted data will result in royalty payments; in this case, system entities can also be rewarded by Buyers (when purchasing data access) or by the system (when inserting unique DNA data).

Alternatively, Collectors may decide to restrict the data access only to a limited number of players. This can be useful when Collectors decide to explore the data themselves, or when Buyers request specific genetic data for their own research. In this case, the proposed platform would still be useful as a Timestamp Server, enabling data owners to register their intellectual property and exchange knowledge.

3.3 Associate Traditional Knowledge

While not strictly necessary for the operation of the Biobank, the system also allows sequenced samples to be associated with traditional knowledge (TK), i.e., data referring to "knowledge, know-how, skill, and practices that are developed and passed on from generation to generation within a community" (WIPO, 2022). Collectors who wish to do so can place the TK in the torrent file associated with DNA data, possibly using different encryption keys, enabling independent control over access to (and monetization of) DNA and TK. All configurations discussed in Section 2.2 also apply to TK, including the registration of smart contracts that define pricing and royalties' rules.

As traditional knowledge might be subject to specific regulations, Amazon Biobank can enforce those requirements through suitable smart contracts. For instance, Brazilian Law N. 13123/2015 foresees cases where the TK benefits must be transferred not only to the Collector as a person, but instead be distributed to an entire community or government. Thus, biocoins received for the rights to access the corresponding TK could be sent to an account registered for the entire community or government. Another example is the need for additional contracts involving free, prior, and informed consent (FPIC). In this case, Amazon Biobank can transparently store those required authorizations, and purchasing a given data implies that the Buyer has accepted those rules. As other rules can be encoded in a flexible way, Amazon Biobank is not restricted to any pre-established model. Nevertheless, we do not claim (or even believe) that Amazon Biobank will be able to automatically ensure compliance with all legislation and with every corner case. This is especially true considering that genetic data may be subject to a multitude of regulations from different countries. However, even in those cases, the Biobank can serve as a useful source of traceability and auditability for transactions involving registered TK, facilitating the resolution of off-system disputes.

3.4 Distribution of DNA Sequences

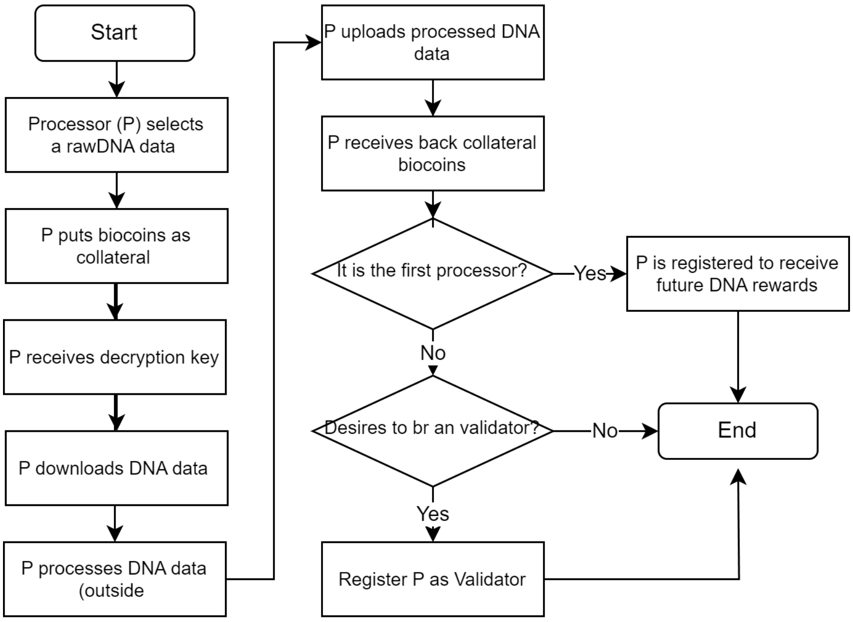
Since raw DNA sequences normally produce large files, a replicated data structure like a blockchain would not be an adequate medium to directly store such large amounts of data. For that reason, as discussed in Section 4.1.2, Biobank uses a BitTorrent-based P2P network (COHEN, 2003) to store and distribute chunks of DNA sequences between different data Distributors. This role can, thus, be played by users with ordinary computers who desire to contribute to the system with their storage and bandwidth capabilities. While doing so, Distributors usually do not gain access to the plaintext contents of those files, unless they purchase the corresponding decryption key using their own biocoins.

Although participation in the Biobank's storage and distribution process is voluntary, Distributors are rewarded with biocoins. Thus, downloaders must pay biocoins for each data piece received by the Distributors. Those downloaders include Processors, which must download raw DNA readings before sequencing data, and Buyers, which download traditional knowledge files and processed DNA data. To ensure fairness and high performance of such transactions, the Biobank system rewards Distributors by using the Torrente blockchain-backed micro-payment mechanism (SHIRAISHI et al., 2021). Note that downloaders can themselves decide to become Distributors after they receive data pieces; in this case, they also are rewarded for each shared piece of data.

3.5 Processing and Validating DNA Sequences

Processors are responsible for executing assembly and sequencing tools on raw DNA data, potentially under different settings, thus creating more useful assembled or annotated DNA sequences. The resulting data and any relevant metadata (e.g., the scripts and programs employed) are turned into a BitTorrent file and added to the system in the form of processed DNA. Processors are then rewarded by participating in the distribution of the benefits for their uploaded DNA data, whether it be royalties or payment to access the corresponding decryption key (Figure 6).

Figure 6 – Flow Chart - Process DNA data


Start
Processor (P) selects a rawDNA data
P puts biocoins as collateral
P receives decryption key
P downloads DNA data
P processes DNA data (outside
P uploads processed DNA data
P receives back collateral biocoins
It is the first processor?
Yes
P is registered to receive future DNA rewards
No
Desires to br an validator?
No
End
Yes
Register P as Validator

Source: Author

However, to obtain the raw data, Processors would have to spend some biocoins for downloading the raw data and for accessing the decryption key. Also, the processing of DNA itself usually takes a considerable amount of processing power. Therefore, Processors must be rewarded in a manner that compensates both the download fees and the subsequent processing tasks. As it is hard to predict the actual interest of a given entry in the Biobank, Processors risk never receiving a return on their investment. The proposed approach for reducing this risk is that for the first few Processors who download the raw DNA data: (1) Federation nodes act as free-of-charge Distributors; and (2) Collectors accept to refund the decryption key fees as soon as the beneficiary Processors register their results in the Blockchain.

Due to the distributed and collaborative nature of the Biobank, more than one Processor can work on the same DNA sequence simultaneously, using the same settings. If not correctly handled, this might lead to very similar assembled or annotated DNA sequences being registered to different Processors. Thus, to avoid undesirable redundancies, the Federation should use a predefined consensus protocol to decide which of the uploaded data is registered as "the first" (and, hence, which Processor is considered the owner of the processed data entry). Nevertheless, the computational efforts of Processors whose results are obtained at a later time are not wasted. Instead, those Processors can assume the role of Validators, confirming or refuting the correctness of previously registered data by similarity with their own results. Together with automated mechanisms (SEPPEY; MANNI; ZDOBNOV, 2019), this approach promotes better data quality and trustworthiness. For example, entries approved by a number of Validators can be considered more trustworthy, so it is fair for those Validators to receive a share of the economic benefits associated with those entries. Conversely, if multiple Validators raise suspicion on some data entry, it can be marked for further scrutiny by Federation nodes. If misbehavior is confirmed, that entry can be removed from the system and its Processor can be punished accordingly. At the same time, those Validators' reputation is incremented, and the data provided by the first Validator who pointed out the misbehavior replaces the previously registered entry.

3.6 Purchasing Access Rights to Data

After finding the raw or annotated DNA sequence of interest, Buyers (usually industry players or academic researchers) can purchase the decryption key for the chosen data, and download it. When that happens, the biocoins spent by the Buyer remunerate all entities involved in the corresponding entry's acquisition and treatment, including Collectors, Processors, Validators, and Curators. This biocoin distribution is defined by smart contracts parameters, as discussed in Section 3.2. Also, those smart contract defines how eventual royalties for products generated from the purchased data should be paid and distributed. Finally, the purchase is registered in the blockchain, which can be used as a transparent log to trace the origin of such biotechnology products to the Amazon Biobank.

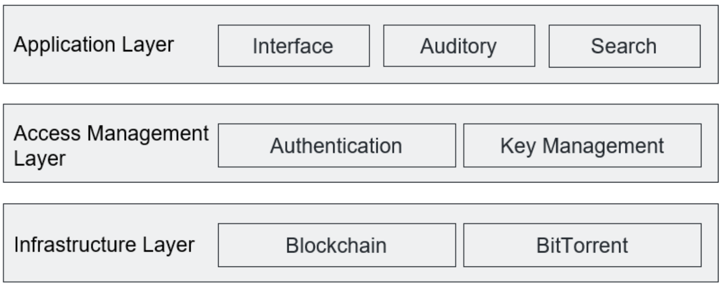
Clearly, neither blockchain nor smart contracts prevent dishonest Buyers from using the acquired data without giving credit to Amazon Biobank or honoring the corresponding payments. Nevertheless, the system's traceability property is expected to be an appealing feature for researchers, who usually need to provide reproducible results, as well as companies having genuine Environmental, Social, and Governance (ESG) policies. Also, data registration Biobank should be useful in case of biopiracy disputes: at the very least, the system clearly shows at which point in time the collected bioresources have been registered and made available for third parties. Hence, even if the corresponding data is acquired via illegal means (e.g., via direct extraction, or in collusion with Buyers who purchased decryption keys), the Biobank can serve as a reference for the prior existence of registered assets and associated metadata.

Instead of periodically searching for data of interest, Buyers can also request data access for a specific organism, remunerating whoever provides the requested DNA sequence. This "on-demand" approach allows system players to focus on data that is in higher demand, increasing their revenues and the value of the Biobank itself. At the same time, Buyers can have their specific needs satisfied more quickly and, possibly, gain priority access to the collected data. Such data access restrictions and corresponding rewards for them are also configured through smart contracts.

4 Proposed Architecture

In this section, we present the architecture of Amazon Biobank. As Figure 7 shows, the architecture can be divided into three layers: an infrastructure layer, an access management layer, and an application layer. The infrastructure layer comprises the core components of Amazon Biobank, the blockchain network, and the BitTorrent network. The access management layer provides identification and authentication services and manages access to protected resources, such as DNA encryption keys. Finally, the application layer provides functionalities that users can directly interact with, comprising the interface applications, the auditing service, and the data search mechanism.

Figure 7 – The architecture layers of Amazon Biobank



Source: Author

4.1 Infrastructure Layer

The infrastructure layer is divided into the blockchain and the BitTorrent network and comprises the main technologies used in Amazon Biobank. The blockchain is responsible for storing the magnet link and all information regarding the genetic data, such as owners, date of insertion, and payments, while the BitTorrent network manages the distribution of the encrypted genetic data.

4.1.1 Blockchain

The Federation maintains the blockchain network, which must register, order, and validate all Amazon Biobank transactions (see Figure 8). For this purpose, each organization from the Federation deploys at least one Endorsing node with the necessary ledger and smart contracts.

Some more engaged organizations may also run an additional Orderer node. As is common in federated architectures, though, the ordering service is less critical than the endorsing service in ensuring the overall trust of the system. The reason is that security issues related to data registration ordering can be handled without relying on the actual trust placed in the ordering service. For example, since nodes are identified, double spending funds can be punished by removing misbehaving peers from the system. As another example, suppose that node N_1 is the legitimate owner of some data whose magnet link is M_1, and, thus, requests the Biobank to register that magnet link under N_1's public key. In this scenario, a malicious node N_2 might try to intercept this request, and then try to register M_1 as its own before N_1 does so. However, such an attack would fail, since N_1 should provide the system with the data corresponding to M_1 only after verifying that the magnet link has been correctly registered in the blockchain. When this check fails, N_1 can simply generate a new magnet link for the same data, in addition to reporting N_2 – whose misconduct can be confirmed after it turns out that N_2 is unable to provide the data corresponding to M_1.

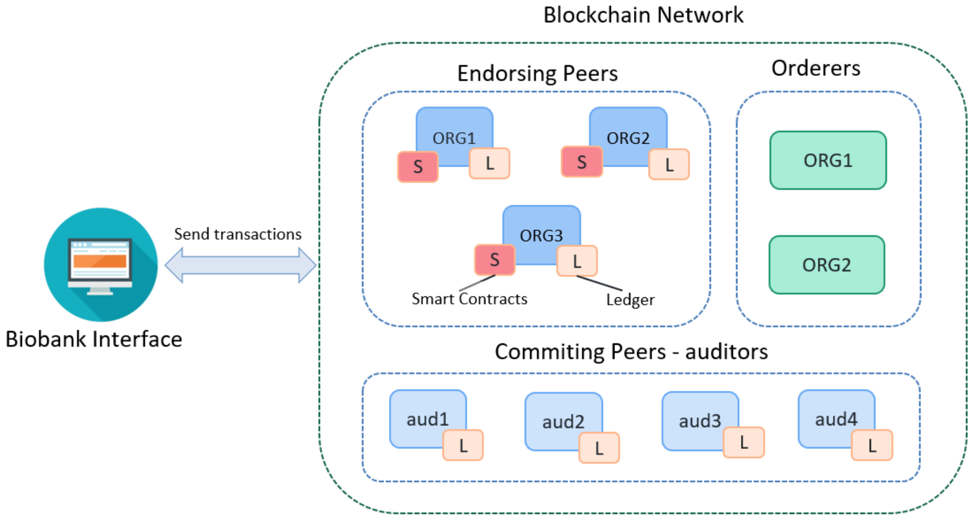
Thanks to this design strategy, the number of organizations running Ordering nodes is expected to remain small (less than ten). Consequently, the ordering service can use a lightweight consensus mechanism without performance concerns. Raft, for instance, can support even 64 nodes without significant performance impacts (GUGGENBERGER et al., 2022).

Finally, the Federation should also deploy some Committing nodes to store the blockchain itself. The Committing role can also be performed complementarily by any number of independent users interested in ensuring the blockchain integrity, without causing a major impact on the performance of the blockchain network.

We have chosen the OR configuration as the endorsement policy of Amazon Biobank, so the action of a single Endorsing node is enough to approve a transaction. As a result, a larger number of organizations can participate in the Federation without affecting the endorsing time. Although this policy might facilitate malicious nodes to approve invalid transactions, this misbehavior can be detected (and, therefore, punished) later, dissuading such attempts. After all, other nodes in the Federation will validate the transaction already published in the blockchain, and apply appropriate sanctions to the identified misbehaving node (see Section 6.1).

The smart contracts deployed in these Endorsing peers contain all the logic necessary for the operation of the Amazon Biobank. In particular, they include tasks related to genetic data (insertion, processing, validation, and purchase) as well as auxiliary functions, such as reputation system procedures. Cryptocurrency-related functions are also included, such as biocoin minting, transferring, and payments for data access.

Figure 8 – Organizations in the blockchain network



Source: Author

4.1.2 Bittorrent

The BitTorrent module coordinates Distributors to store and share encrypted DNA sequences. To do so, it first employs a centralized Federation server that stores a backup of all genomic data. Then, to increase data redundancy, the module requests Distributors to download and seed some of those genomic data. A BitTorrent tracker monitors the number of Distributors seeding a given genomic data. Thus, if this number decreases excessively, the module requests more Distributors to assist with this genomic data distribution. The help of Distributors is also positive for network scalability: as more Distributors are seeding a given data, the greater the network bandwidth capability.

In Amazon Biobank, downloaders must pay for each piece of data with biocoins. This applies even to the Distributors; thus, to start seeding, Distributors must first invest some biocoins. Therefore, Amazon Biobank adopts an alternative incentive mechanism to encourage new Distributors. Every time new data is inserted into the system, volunteer Distributors receive special freshly minted biocoins. Those biocoins can only be used to download the new genetic data from the Federation, and if not used in a given time, they are reassigned to other Distributors. Thus, new Distributors have the opportunity to start seeding for free, and Amazon Biobank can more easily increase its data redundancy.

4.2 Access Management

The Access Management layer manages the permissions related to Amazon Biobank's data access. The Authentication module is responsible for user creation and authentication, and the Key Management module controls the access to data decryption keys.

4.2.1 Authentication

In Amazon Biobank, user creation is managed by each organization via Certificate Authorities (CAs). To do so, each organization runs its own CA and validates new user accounts through their own identification methods (e.g., password authentication, manual verification of the person's ID, OAuth integrated with corporate email, or biometry). Other organizations may verify this process indirectly: if they are detected too many fraudulent user accounts (e.g., fake or duplicated accounts), the issuer organization may be punished (e.g., by being evicted from the system).

A CA can create a user account by issuing a credential file, which includes a certificate and the corresponding private key. This credential is inserted on the Biobank interface to sign all transactions sent to the blockchain. To avoid inadequate access to this credential, users may encrypt the credential via password-hashing techniques. In addition, this credential never is transmitted to any remote non-trusted environment; thus, only the biobank interface, executed locally, can access the credential. The biobank interface can be open source to avoid any misbehavior suspicion.

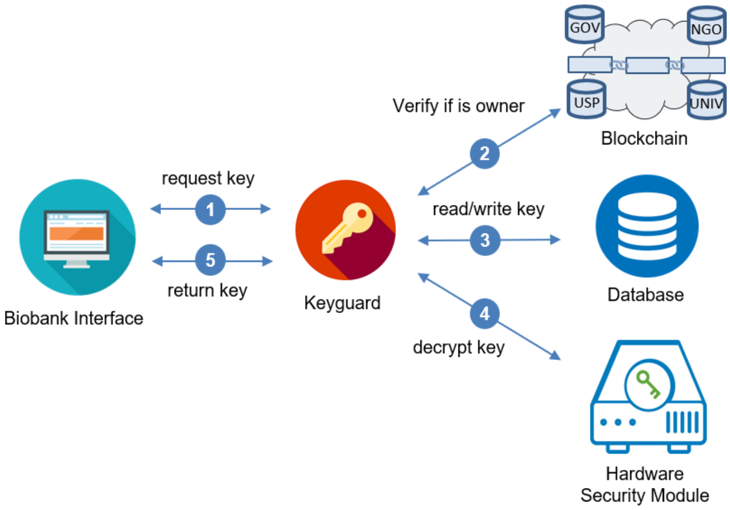
4.2.2 Key Management

The key management module, or Keyguard, is a module maintained by the Federation and manages access to data decryption keys. To do so, Keyguard verifies the credential used to sign the transaction, identifies the requesting user, and compares it with the data owners registered on the blockchain. Fig. 9 illustrates the steps necessary to access the decryption key. Note that a Hardware Security Module (HSM) can be used to encrypt all decryption keys before storing them in a database, thus protecting them even better against data breaches. Also, secret sharing schemes can be employed to distribute the key management between different entities of the Federation, avoiding a single point of failure. Nevertheless, some Collectors may prefer to keep decryption keys to themselves. In this case, they handle their own key management service, together with the search and/or backup services usually provided by the Federation.

4.3 Application

The application layer includes the Amazon Biobank modules through which users can interact with the system. Those modules are (1) the user interface, where the user can insert, download, and purchase access to genetic data; (2) the auditing module, which allows users to verify the correct operation of the biobank; and (3) the search management.

Figure 9 – Key management steps to access a data decryption key



Source: Author

4.3.1 Interface

The Biobank interface is responsible for user interaction with the Amazon Biobank, allowing data reading and writing. It can interact with the blockchain to insert, download, and purchase access to genetic data; and to define some smart contract parameters, such as royalties and price. It can also interact with BitTorrent to store and distribute DNA data, encrypt and decrypt DNA sequences, and generate .torrent files and magnet links. In addition, the interface also implements the Torrente micropayment protocol, paying and requiring biocoins for each piece of data transmitted. Finally, the interface also facilitates the use of third-party applications (JAIN et al., 2016) to process and validate inserted DNA sequences.

4.3.2 Auditing

The blockchain auditing module allows any interested user to monitor and validate Amazon Biobank's correct operation. To do so, independent auditors have three options to access blockchain data. First, they can access the blockchain via API calls. The Federation runs an API service that provides transactions, blocks, and any data of interest by directly communicating with the blockchain. Second, auditors can use the blockchain viewer, a data-friendly web interface. Amazon Biobank uses Hyperledger Explorer (HYPERLEDGER. . . , 2022b), a default visualizer tool for Hyperledger Fabric, and provides a dashboard similar to those present in many cryptocurrencies (e.g., Blockchain.info[1] or Etherscan[2]). Finally, auditors can store and monitor the blockchain's blocks through IPFS. IPFS provides an alternative and distributed way to store the last block of the blockchain to prevent alternative history attacks (see Section 6.1). To do so, the Federation periodically publishes

[1] <https://www.blockchain.com/explorer>

[2] <https://etherscan.io/>

on IPFS the blockchain's latest block, and auditors access it to compare with the version published through API. If desired, auditors can run their own IPFS nodes to participate in the network directly, monitoring and preventing a block substitution.

4.3.3 Search Management

Users can employ different options of search to find genetic data of interest. The most usual is the (1) traditional keyword-based searches, done over the associated metadata. In this, users insert some keywords of interest, and the Federation scans all stored metadata looking for matching results. If a decentralized search is desired, those metadata and the corresponding keywords can also be stored in a DHT. Another possibility is the (2) encrypted keyword search, in which searchable-encryption mechanisms can be used to enable users to query data of interest privately.

Finally, in (3) sequence-based search, users insert a sequence of interest, and the system returns all matching genomic data. To do so, data owners may share the decryption key with the Federation, which analyzes the DNA sequences stored in its backup. As this search requires high computational costs, users might need to pay a steeper fee to request this operation.

The sequence-search costs can be reduced with the collaboration of other nodes. For instance, players with access to genomic data (e.g., Processors or Buyers who previously acquired the DNA sequences) can analyze it, return the result to the Federation, and receive part of the corresponding search fees. Thus, Collectors, Processors, and Validators not only receive research fees but also increase the visibility and the selling potential of their data. The Federation must validate each result before returning it to the user, penalizing any incorrect result (e.g., by decreasing reputation). Nevertheless, Collectors may choose not to share the decryption key with the Federation, nor allow collaborative searches. In this case, Collectors are responsible for handling their own data search.

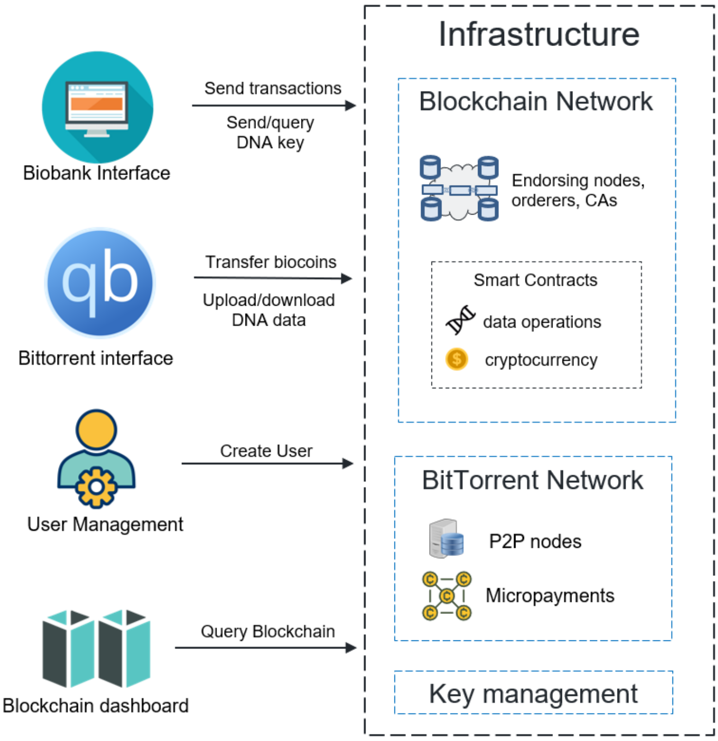
5 Results

In this section, we describe the implementation of Amazon Biobank and the experiments conducted to evaluate the system's performance.

5.1 Prototype Implementation

To illustrate the operation of Amazon Biobank, we built a prototype with blockchain, BitTorrent, and other layers as described in Section 4. Consequently, our prototype allows main operations over genetic data, including data registration, purchase, and download. Figure 10 shows some of the implemented modules. The prototype and the corresponding documentation are available in <https://github.com/amazon-biobank/biobank>, and a step-by-step tutorial for some operations is described in Appendix A.

Figure 10 – Main modules of the Amazon Biobank prototype



Source: Author

For the blockchain layer, we used Hyperledger Fabric to deploy a demo network with three endorsing nodes and one orderer. On each endorsing node, we developed and deployed smart contracts that implement the core functionalities of Amazon biobank. Table 2 shows the prototype's main smart contracts. For instance, the DataContract manages the registration of the raw genetic data, inserted by the Collector. This contract ensures,

Table 2 – Main smart contracts from Amazon Biobank prototype

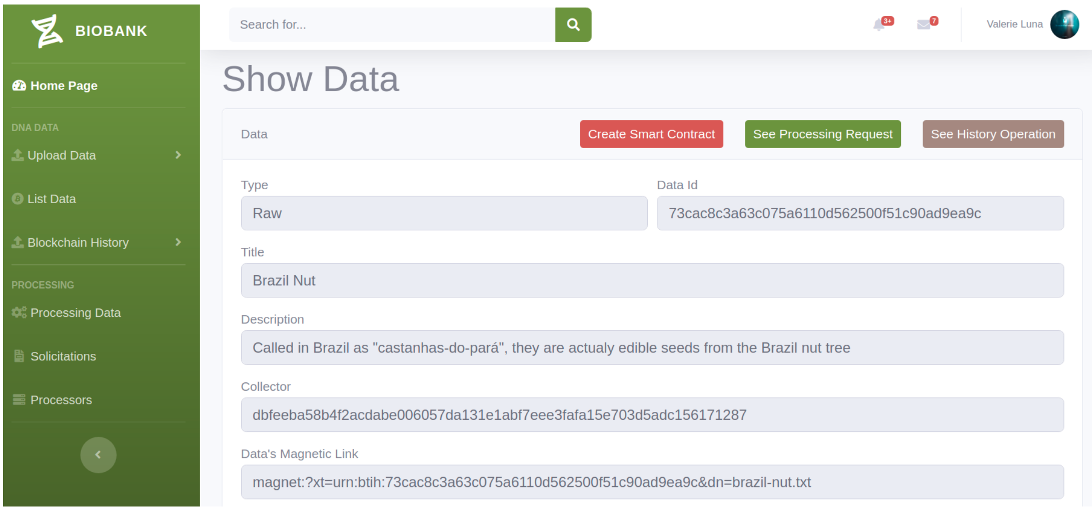
Smart Contract	Description
Data	insert and query raw and processed data
DNAContract	enforces the selling conditions (e.g., price and payment distribution)
ProcessRequest	coordinates the processing tasks
Account	manages the user account
Biocoin	implements the biocoin transference and payments

Source: Author

for example, that the magnet link is uniquely registered on the blockchain, preventing duplication.

To interact with the blockchain layer, we built the biobank interface, an express.js web application. Users employ this interface to register new genetic data, purchase a DNA decryption key, or make any other operations on the system (Figure 11). To allow this, the interface interacts with the blockchain layer by sending signed transactions. The communication between the interface and the blockchain network is done through Hyperledger Fabric SDK[1]. The interface also implements password hashing, thus users can use encrypted credentials to authenticate themselves, as described in Section 4.2.1.

Figure 11 – Accessing DNA sequence information through biobank interface



Show Data

Data

Type: Raw

Data Id: 73cac8c3a63c075a6110d562500f51c90ad9ea9c

Title: Brazil Nut

Description: Called in Brazil as "castanhas-do-pará", they are actualy edible seeds from the Brazil nut tree

Collector: dbfeeba58b4f2acdabe006057da131e1abf7eee3fafa15e703d5adc156171287

Data's Magnetic Link: magnet:?xt=urn:btih:73cac8c3a63c075a6110d562500f51c90ad9ea9c&dn=brazil-nut.txt

Source: Author

For the BitTorrent layer, we used Torrent (SHIRAISHI et al., 2021), a dedicated client application based on qBittorrent[2]. By installing this application, users can turn their machines into a BitTorrent node, contributing to the P2P genetic data distribution. The Torrente application also implements the micropayment mechanism, which requires a biocoin payment for each piece of data transmitted. To do so, it communicates with the blockchain layers, to redeem biocoins for a given account. The Torrente application also

[1] https://hyperledger.github.io/fabric-sdk-node/

[2] <https://www.qbittorrent.org>

includes a user interface, by which users can activate the commands to upload/download a given genetic data. In addition, the Torrente interface supports operations auxiliary to data distribution, such as encrypting new genetic data through AES-GCM and creating the ".torrent" file as well as its corresponding magnet link.

We also built a preliminary version of the modules related to access management, in Amazon Biobank. Our Key Management module is responsible for storing the DNA decryption keys and was built using express.js and MySQL. It also can communicate directly with the blockchain to verify the permission of the requesting user before allowing access to data decryption keys. In turn, we built a user management system that simulates a user management control that an organization would implement. In our user management system, the WEB application uses oAuth to verify the user's identity, and it authorizes new user accounts only for accepted domains (e.g., @usp.br and @unir.br domains). Then, it communicates directly with the organization's CA to create new user accounts.

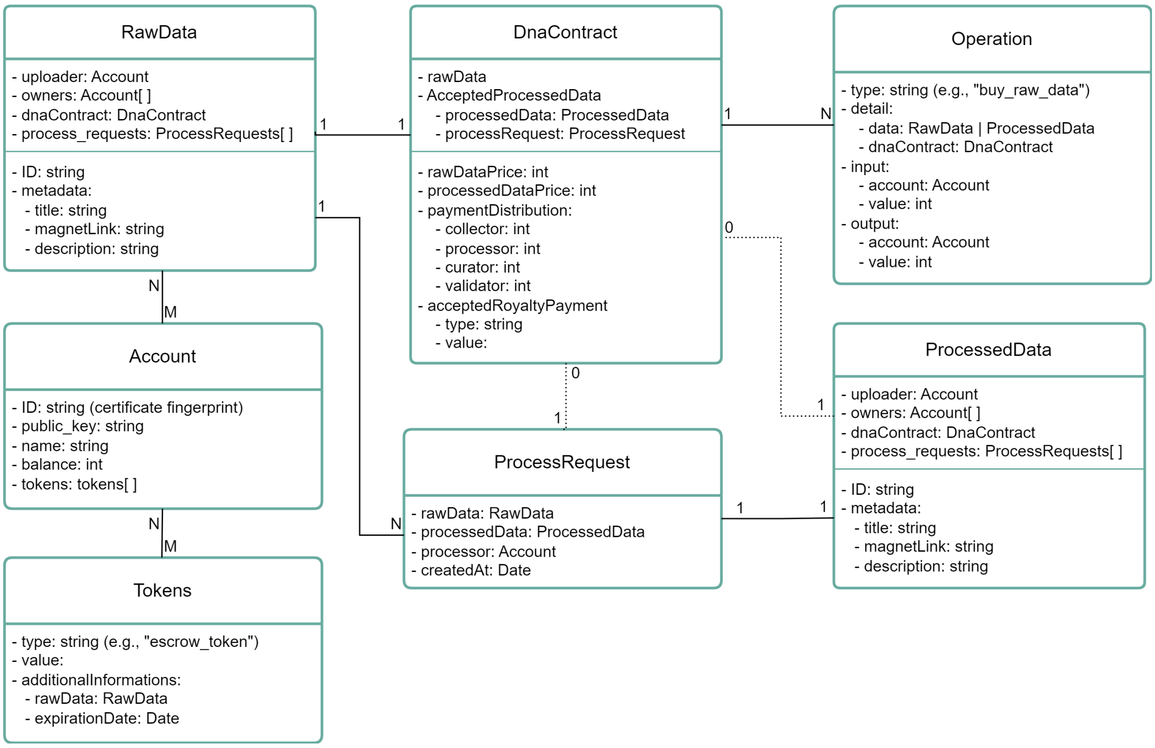
Finally, we deployed a blockchain visualizer tool through Hyperledger Explorer (DHILLON; METCALF; HOOPER, 2017). This visualizer facilitates blockchain auditing by allowing access to blockchain data, including transactions that were sent to the system, blocks that were approved, and the number of peers and orderers present on the network. As this dashboard does not require authentication, it can be accessed by any interested user.

Figure 12 shows the main objects that are registered on the blockchain through Hyperledger's smart contracts. In the diagram, the rawData represents the raw genetical data that can be inserted by the Collector. In turn, the DnaContract contains some parameters for a Buyer to purchase access to DNA data (e.g., price and payment distribution). The processingRequest object represents the Processor's assignment for a given rawData, and the processedData corresponds to the data resulting from this processing. Then, when genetic data is purchased, the Operation object registers the transaction details such as value, involved users, and the conditions present on the dnaContract. Finally, Account represents each user on the system, keeping track of their biocoins balance. Those accounts may also have some tokens, which are auxiliary objects for payment operations, such as the redemption of micropayment for BitTorrent.

5.2 Performance evaluation

In this section, we evaluate the performance of the Amazon Biobank. We first tested several blockchain configurations, varying the number of nodes, organizations, and computational resources. Then we measured the download time of BitTorrent with and without the Torrente micropayment and compared it to centralized servers.

Figure 12 – Overview of the entity relationship diagram



Source: Author

5.2.1 Blockchain performance

To conduct experiments for the blockchain, we deployed each Hyperledger Fabric endorsing node in a lightweight Ubuntu virtual machine, with 1vCPU (2.3 GHz) and 2GB of RAM. Each of those nodes received the prototype smart contracts, such as dataContract and processRequestContract described in section 5.1.

We then measured the latency and the throughput of two types of transactions. The first type is a data write transaction, representing the registration of new genetic data. This transaction inserts a new rawData object with the corresponding metadata and magnet link on the blockchain. The second type is a data read transaction, representing a user inspecting details about a given DNA. It queries a rawData on the Hyperledger Fabric ledger through the data ID. To generate the transaction workload, Hyperledger Caliper was used (HYPERLEDGER. . . , 2022a), a performance benchmark tool specialized in Hyperledger Fabric. In each of the tested scenarios, the amount of failed transactions was insignificant (less than 0.01%). Other details can be seen in Table 3.

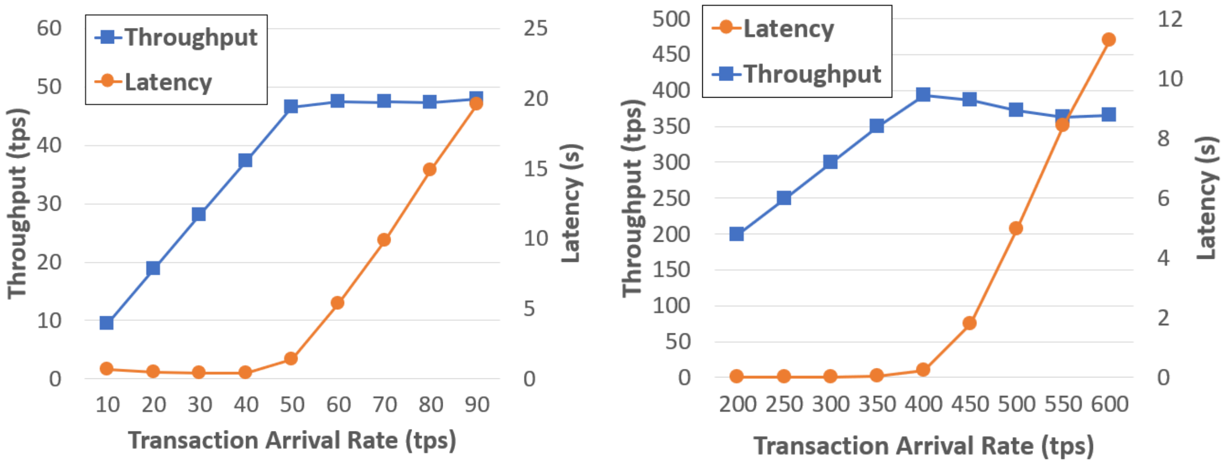
We first tested the prototype in a configuration that simulates the initial deployment of Amazon Biobank. Thus, we deployed 3 organizations with 1 endorsing peer each, representing a Federation composed of a few universities or NGO institutions. Our experiments indicate that the prototype supports at least 50 tps of data write and 400 tps of data read (Figure 13). As we expect limited use of the system in the initial phases, this performance can be considered satisfactory.

Table 3 – Default configurations for all experiments unless specified otherwise

Parameters	Values
Number of Channels	1
StateDB Database	CouchDB
Network configuration	2 organizations with 1 peer each and 1 orderer
Peer Resources	1vCPU (2.3 GHz), 2GB RAM
Endorsement Policy	OR
Block Size	500 transactions per block
Block Timeout	2
Number of orderers	3

Source: Author

Figure 13 – Prototype's base performance (3 organizations with 1 peer each)



(a) Data write

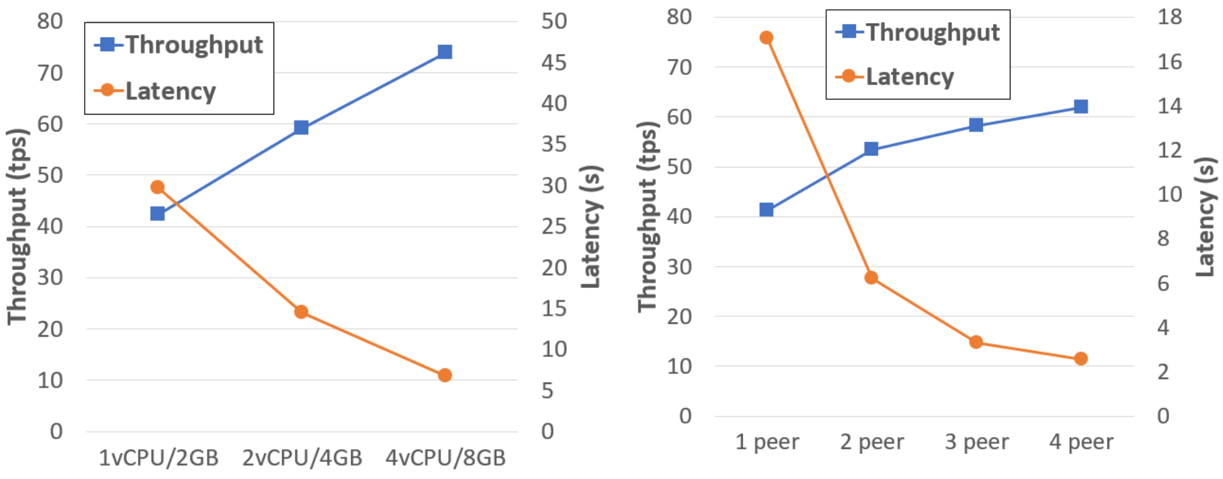
(b) Data read

Source: Author

Still, Amazon Biobank might require more performance as gains popularity; thus, we investigated some alternatives to improve it. One possibility is to upgrade the computational resource for each node. As Figure 14a shows, an upgrade from 1 vCPU to 2VCPU can increase the transaction rate by 50% while cutting the latency time by less than half. Another option is to increase the number of Endorsing peers in each organization. Thus, we deployed additional endorsing nodes for each of the three organizations defined in the initial configuration. As Figure 14b indicates, the addition of even one extra peer can cut the latency time by less than half and increase the throughput by almost 50%. Those results indicate at least two options for the Federation to improve the blockchain performance if required.

With the popularity of Amazon Biobank, more organizations are expected to contribute to the system. Thus, we also investigated the effect of adding new organizations on the Federation. It is possible that an increase in the Federation size could result in a longer consensus time, affecting the performance. Interestingly, our experiments indicate the opposite: blockchain performance is rather improved with more organizations

Figure 14 – Gains on performance by (a) improving computational resources from each node; or by (b) increasing the number of peers in each organization. Tests were done using data write transactions, and the arrival rate was 100 tps and 70 tps, respectively.



(a) Computational resources

(b) Number of peers on each organization

Source: Author

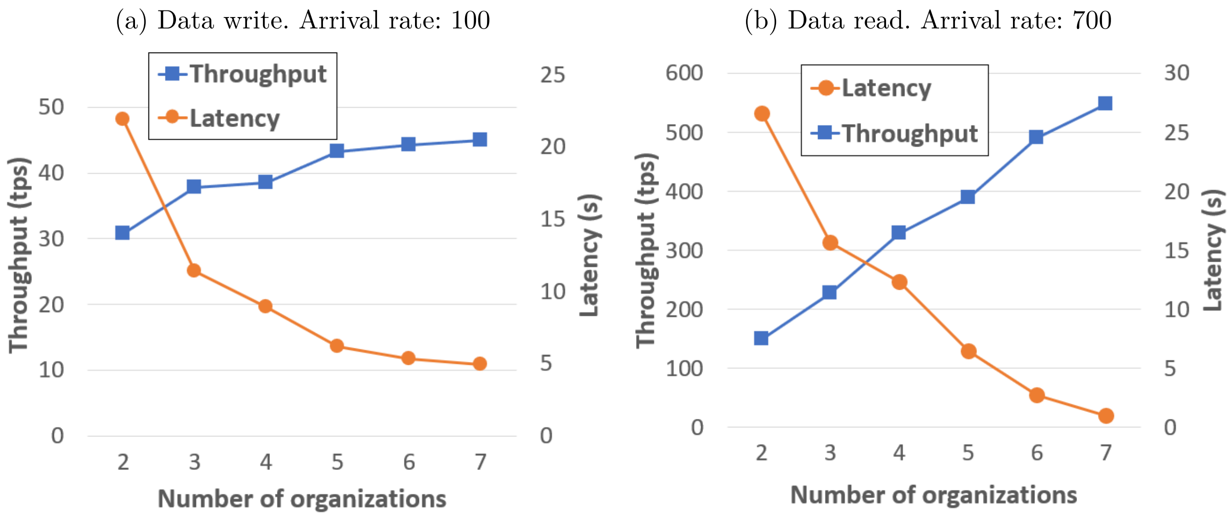
in the Federation. Figure 15a shows this behavior: the data write transaction rate is slightly improved when new peers are added. This can be explained by the adoption of a permissive endorsing policy (OR), which keeps constant the number of necessary endorsing nodes' validation (Section 2.2). Similarly, Figure 15b presents a linear improvement in the data query transactions rate. In this case, the query workload is divided between more organizations, alleviating the load pressure in each node.

For each of those experiments, the number of three ordering nodes was kept constant. This was chosen because, as mentioned in Section 2.2, the ordering phase is less critical than the endorsing phase for the system's trust. Thus, we expect only a few organizations to run an orderer node. In this scenario, some studies show that adding new ordering nodes does not cause bottlenecks in the system (GUGGENBERGER et al., 2022). Besides, in Amazon Biobank, new organizations must necessarily deploy an extra endorsing node. Therefore, even if an organization deploys an additional orderer, the extra endorsing node will most likely result in a performance increase (Figure 15).

5.2.2 BitTorrent performance

For the experiments, BitTorrent nodes were deployed in personal computers with Intel core i5 and i7 and 8GB RAM or more, and those nodes were connected to a local network. Then, to establish a common network resource, the upload rate of each node was limited to 1MiB/s. A file with random data was used for distribution, and it was transmitted through pieces of 16 KiB size.

Figure 15 – Influence of the number of organizations on performance.


(a) Data write. Arrival rate: 100

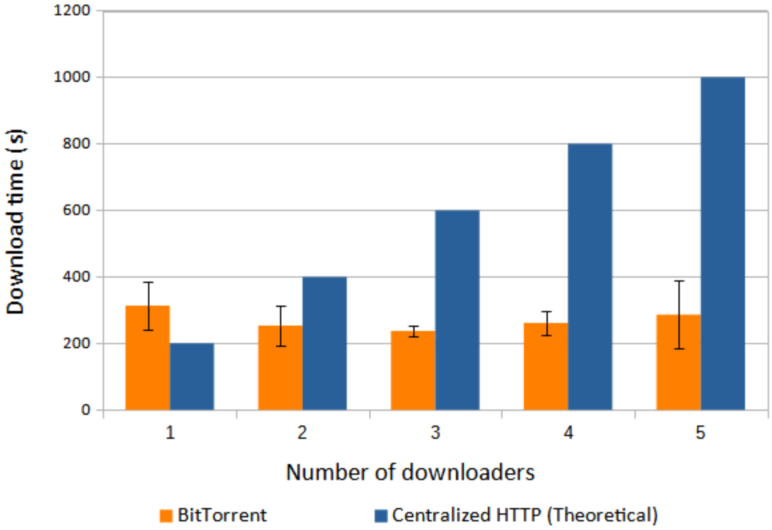
(b) Data read. Arrival rate: 700

Source: Author

For the first experiment, we represented a Collector inserting new genetic data into the system. In this scenario, the Collector acts as a data seeder, and the Distributors are the downloaders acquiring the data. Thus, we configured 1 seeder with a 200MB file and 1-5 simultaneous downloaders. Figure 16 shows the download times. Note that, in the BitTorrent scenario, the download time tends to maintain constant even with an increasing number of downloaders. Thus, BitTorrent is suitable for big-size files (typical of genetic data) or a great number of simultaneous downloaders. In contrast, the download time for a centralized HTTP server (limited to 1 MiB/s bandwidth) only increases with the number of downloaders. This difference can be explained by the P2P file sharing mechanism: in BitTorrent, downloaders obtain data not only from the original seeder but also from other downloaders. On the other hand, in a centralized scenario, downloads receive data only from the data server, often overloaded.

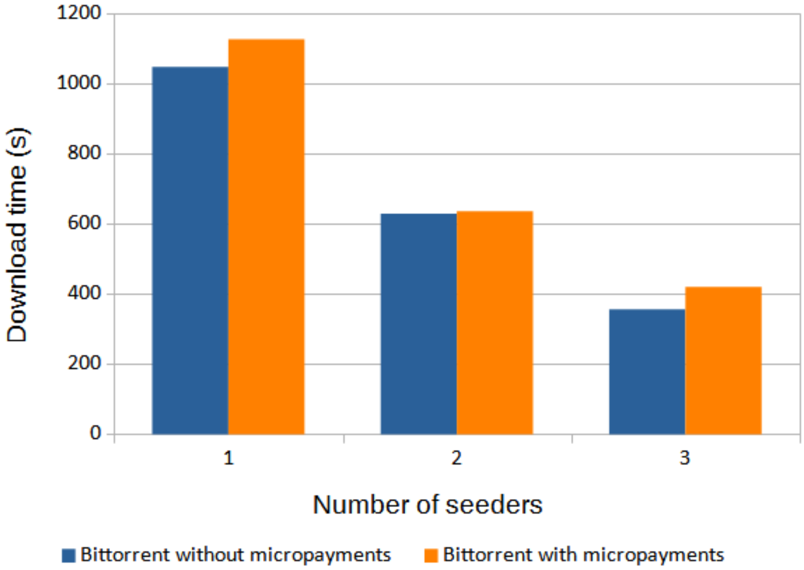
Then, we evaluated the performance of the Torrente micropayments. Specifically, we measured the overhead introduced by this mechanism by comparing the download time with and without micropayments. To do so, a 1 GB size file was created, and the download time with 1,2, and 3 seeders was obtained. Our experiments show that the micropayment increases the download time by 10%, on average (Figure 17). This overhead is significantly lower than the gain obtained by the addition of one extra seeder (near 30% gain). Therefore, the benefits of financially encouraging new seeders can be considered enough to compensate for the micropayment overhead.

Figure 16 – Time taken to download a 200 MB file via BitTorrent and via centralized server



Source: Author

Figure 17 – Comparison between BitTorrent download time with and without micropayments



Source: Author

6 Discussion

Amazon Biobank promotes the collaborative development of biodiversity-based research in regions with rich ecosystems. By combining blockchain, smart contract, and P2P technologies, the system fulfills the requirements described in Section **??**:

- The system allows Collectors to upload raw DNA data and the correspondent metadata into the Biobank. Processors with enough computational capabilities can then download and process those raw sequences, uploading annotated sequences. The system's data storage and bandwidth capabilities are reinforced by a distributed P2P network formed by Distributors. Those players contributing with computational resources are rewarded with biocoins, the system's internal currency. Biocoins can then be converted into fiat currencies via Exchanges, which can also handle direct trades among interested parties (e.g., purchases of biocoins by Buyers who want to access the system's genetic data). The larger the amount of data inserted into the system, the more it should attract the attention of potential Buyers, increasing the value of biocoins according to supply and demand rules.

- The system registers all operations related to a DNA sequence, including raw data collection, processing, distribution, and purchase of access rights. That way, in case of legal disputes, users can use the transparent log provided by the underlying blockchain as proof of such events. This approach is expected to facilitate real-world actions regarding intellectual property protection.

- Curators and Validators help to promote data and metadata quality, increasing confidence in the accuracy of the corresponding entries. If misconduct is detected, the culprit's reputation is penalized; in case of repeated misbehavior, the user may be evicted from the system, losing access to existing funds and to future profit opportunities.

- The system is organized in such a manner that Federation nodes can perform searches in local data, as well as collaborate with other nodes (e.g., Collectors, Processors, and Buyers) for that purpose.

- To legitimately access some DNA sequence, Buyers must invest some biocoins. Except for eventual system fees, the total amount paid by Buyers is then shared among all players responsible for the availability of that data entry (not only Collectors but also Processors, Distributors, Validators, and Curators). If some profit is made thanks to that data, a certain amount of royalties is also expected to be reverted to those players, according to the rules established in smart contracts configured by the data stakeholders. Buyers can also enable public access to genetic data by paying the correspondent fees to the Collector and other relevant actors; similar to the open-access principle for academic publishing.

Similarly, the system also fulfills its non-functional requirements:

- Auditability and Traceability: all data is converted into torrent files for integrity protection, and operations are recorded in the blockchain to create a temporal and transparent log of events. Hence, any research or product developed from the registered data can be securely traced back to the corresponding entry in the Biobank. For example, this allows Buyers to prove the relationship between their products and Amazonia's biodiversity. Once some data entry is registered in the system, attempts of modifying its contents or place in time can be detected by auditors. Anyone can audit the blockchain and verify the correctness of operations thereby registered, so the system's transparency is independent of the amount of trust bestowed upon the Federation itself.
- Scalability: the system leverages collaborative distributed technologies, like BitTorrent, to avoid many of the scalability issues typically found in centralized platforms (e.g., bandwidth and storage limitations). In addition, the burden of executing computationally intensive tasks, like DNA processing and sequence-search operations, can be shared among system players. The resulting collaborative architecture is, thus, expected to be capable of handling a large number of users and data.

6.1 Blockchain Correctness

Amazon Biobank can show the correct operation of the system even in the presence of malicious system operations. To do so, any member of the Federation, as well as any independent auditors, monitors the blockchain and verifies that no past transaction has been modified. Since every node maintains a copy of the blockchain content, even one honest node is enough to detect manipulation attempts. In other words, the blockchain can only be silently modified (i.e., without being detected or raising any alarm) if every node in the network consensually decides to do so.

Therefore, Amazon Biobank is open only to a "100% attack", in which every node from the network needs to coordinate to overwrite past transactions. Note that this approach gives stronger integrity guarantees than some popular public blockchains. Bitcoin, for example, adopts a "longest chain" rule to decide which should be considered the valid version of the blockchain. This means that nodes silently overwrite their current

Table 4 – Some of the considered "100% attacks" and its detection

Attack	Description	Detection
Discrete modification	Modify a past transaction without updating the following blocks	Verify all blockchain consistency
Alternative history	Overwrite the blockchain history with a manipulated chain	Monitor blockchain checkpoints
Inadequate endorsing	Endorsing peers approve an invalid transaction	Verify transactions on the blockchain

Source: Author

blockchain version if they receive a different version containing more blocks. As a result, any entity (or group of entities) that can create blocks faster than its peers can overwrite past transactions at any moment, (e.g., 51% attack or selfish mining attack)(EYAL; SIRER, 2018). In comparison, Amazon Biobank does not accept any such “longest chain” rule. As Hyperledger Fabric is deterministic, every time a block is published, it should never be overwritten by another blockchain version. Therefore, if two versions of the blockchain exist, this necessarily means that some Federation node has generated the second version maliciously.

To illustrate the system’s resistance, we describe how auditors could detect attacks involving even 100% of the Federation. Table 4 shows some of those attacks.

In the first attack, Federation nodes make a discrete modification in a past transaction without updating the following blocks. This results in an inconsistent blockchain, in which the hash of the modified block does not correspond with the following blocks. Thus, independent auditors can detect it when verifying all the hashes and signatures from the blockchain. To help with this task, auditors can use automatic verification tools (e.g., BCVerifier (SHIMOSAWA; SATO; OSHIMA, 2020)).

A second attack is the alternative history attack, in which the Federation nodes overwrite the last blocks of the blockchain with a modified version, presenting it as the correct one (DABHOLKAR; SARASWAT, 2019). To detect this, auditors must periodically store and monitor the latest block of the blockchain as a checkpoint, ensuring that it is not substituted. If, during the verification, auditors realize that a checkpoint has been removed from the blockchain, this fact can be used as proof that the Federation has misbehaved by substituting the blockchain contents.

To facilitate access to those checkpoints, the Federation publishes them in a public and easy-to-audit data storage, as explained in section 4.3.2. Usually, those checkpoints are published in public blockchains (ROBINSON; BRAINARD, 2019); however, this results in higher costs due to transactional fees. Therefore, Amazon Biobank adopts a someway cheaper alternative, IPFS. Periodically, the Federation publishes the checkpoint in a self-signed domain in IPFS, and auditors can request it for monitoring. Due to distributed nature of IPFS, auditors can even run their own IPFS nodes to prevent data modification from the Federation. In addition, the Federation can also publish the world state (the aggregate result after executing all transactions from the blockchain up to that checkpoint). Thus, unsatisfied auditors can start their version of Amazon Biobank at any time by effectively “forking” the blockchain from the last reliable world state.

Finally, a third possible attack is when the Federation runs malicious endorsing peers to approve invalid transactions. As explained in Section 2.2, if enough endorsing peers collude, they might register transactions that do not comply with the chaincodes deployed on the network. To avoid that, independent auditors periodically verify the transactions registered on the blockchain. They can access the details of each transaction, re-execute them on the chaincodes and check if the generated result is equal to the registered result. If any inconsistency is detected, endorsing peers that signed those transactions are investigated and penalized.

6.2 Remaining Challenges

Although Amazon Biobank addresses many of the concerns faced by collaborative genetic repositories, some challenges still remain. Those issues are similar to those discussed in other blockchain-based management systems, and solutions have not yet been definitively established (ITO; O’DAIR, 2019). Amazon Biobank provides mechanisms to attenuate, rather than completely solve, those problems.

One significant challenge is how to prevent the insertion of false or forged DNA data. While automatic detection mechanisms exist (SEPPEY; MANNI; ZDOBNOV, 2019), it can be difficult to identify data that has been specifically forged to bypass these mechanisms. To mitigate this challenge, Amazon Biobank employs a data verification process that involves Curators, Processors, and Data Buyers verifying the genomic data after it has been inserted. Collectors are rewarded only after their data are verified, and users who engage in misbehavior may face punishments, such as loss of reputation or suspension from the system.

A second challenge is that Amazon Biobank cannot prevent DNA data distribution outside of the system. For example, a user with access to the decryption key may decide to share it freely with others. However, we argue that users have strong incentives to use Amazon Biobank due to the traceability and auditability benefits it provides. For instance, companies with environmental, social, and governance (ESG) policies usually need to produce evidence of their efforts, and researchers need to provide reproducible results for their publications. In addition, sharing the data outside the blockchain only hinders the user, as it increases the competition without resulting in any concrete gain.

To try to prevent this out-of-the-system data distribution, Amazon Biobank may use watermarks. That way, unauthorized data and the users who leaked it could be identified and punished. For example, some studies propose hiding those watermarks in non-critical parts of the DNA (NA, 2020). This could dissuade at least less sophisticated users from sharing genomic data. However, users still could remove the watermarks from the genomic data or rewrite them with random values. The location of the watermarks could be made secret, but this "security by obscurity" approach fails the instant that their operation needs to be revealed (e.g., to prove the watermark's existence in a judicial court). Thus, watermarks do not offer a definitive solution to this challenge.

Deploying the blockchain network within the Federation has also its own set of challenges. In particular, special care must be taken to protect against DOS attacks (BATTISTI et al., 2022). As Amazon Biobank relies on a permissioned blockchain, a malicious attacker could send a massive amount of transactions to a specific node, exhausting its computing and memory resources, and provoking failed transactions. In addition, the criterion to include a given organization in the Federation must also be defined. Otherwise, a malicious attacker could create one or many organizations to attack the system, harming the blockchain stability.

Besides those technical challenges, Amazon Biobank also needs to consider local and international regulations. Typically, the government has rules to regulate biodiversity exploration, varying from notification to the authorities to royalty payments. In addition, the monetary value of biocoins might require some adjustments to enforce financial compliance. Collaboration with other stakeholders, such as governmental agencies, cryptocurrency exchanges, and potential users of data, might be required.

Finally, the proposed system is designed to simply handle biodiversity data, making no claims over copyright or restricting the usage of that data. Specifically, we explicitly exclude human and medical samples from the scope of the Amazon Biobank. After all, supporting this kind of data in a collaborative system would require additional legal and ethical considerations. Similar legal challenges may arise from handling associated traditional knowledge (TK). Nevertheless, we envision that registering TK in an encrypted form may be interesting for traditional knowledge holders, even though it is not strictly necessary for the functioning of the Biobank. At least, handling TK can be useful to claim ownership of that knowledge at the time of registration, if that ever becomes a matter of dispute (e.g., by other groups that share the same TK). Eventual access to TK, if allowed, can be mediated through blockchain and smart contracts. In practice, however, any implementation supporting registration and access to TK would necessarily involve collaboration with stakeholders, including indigenous populations, governmental agencies, and potential users of data.

Table 1 – Comparison between blockchain-related genetic projects

	Data Validation	Distributed processing	Sequence Search	Benefit Sharing	Association with owner
Encryptgen	●	-	-	-	-
Zenome	●	●	-	-	-
Nebula Genomics	●	●	●	-	-
Genesy	●	●	●	-	-
Global ABS Tracker	-	-	-	●	●
Amazon Biobank	●	●	●	●	●

●=provides property; -=does not provide property;

Source: Author

6.3 Related Works

As briefly mentioned in Section 1.1, many blockchain-based genomic repositories have been proposed to remove intermediates and increase user control over their data. Table 1 presents some examples of those projects, comparing them based on features relevant to biodiversity research.

One of the proposed genomic marketplaces is Encryptgen (ENCRYPGEN, 2017), a platform in which users can provide their genetic data in exchange for cryptocurrency tokens. Similarly to Amazon Biobank, it stores DNA data in an encrypted form and registers the metadata and the transactions on the blockchain. That way, users can choose when and to whom to give access to their genome. However, the platform has little support to distribute genome processing tasks collaboratively. For instance, users cannot request others to assist in genomic sequencing and assembling, an essential step after collecting raw DNA data. In addition, Encryptgen does not allow genomic analysis directly on uploaded data without downloading it. Consequently, it does not support sequence-based searches (finding genomic data based on a given DNA sequence).

In contrast, the Zenome platform (KULEMIN; POPOV; GORBACHEV, 2017) presents the role of a computational node, a participant that provides storage and CPU time in exchange for ZNA tokens. Thus, similarly to our proposal, the processing and the sharing tasks can be distributed to other participants, decentralizing the data storage and analysis costs. In addition, Zenome employs a data rating system, in which high-quality data are identified while less valuable data are penalized. Nevertheless, the Zenome platform also does not support sequence search operations. Although it is possible to perform queries to the system, extracting statistical data regarding genomic status or user questionnaires, those queries are supported only over unencrypted data. Therefore, users must insert the data either in (1) unencrypted format, abandoning the data privacy and confidentiality; or in (2) encrypted and no-searchable format, renouncing the data visibility provided by sequence searches.

In turn, Nebula Genomics (GRISHIN et al., 2018) allows computation over encrypted DNA data by using privacy-preserving techniques, such as partially homomorphic encryption and secure computation over Intel Secure Guard Extension (SGX). Therefore, it can better conciliate data privacy with distributed data processing operations. Additionally, Nebula Genomics aims to transfer the costs of genomic sequencing to data end-users. Thus, interest third-party companies, such as pharmacies and research centers, may subside sequencing costs for users with unusual phenotypes. This reduces the entry costs for data sequencing. Nevertheless, this business model decreases the user's control over genomic data: once the data is inserted, the user has little control over it. In contrast, Amazon Biobank allows users to better control their genomic data through configurable smart contracts, defining the price and the conditions for the data utilization.

Genesy (CARLINI et al., 2019) is another genomic marketplace that provides sequencing services, such as genomic data access and the distributed sharing of DNA sequences. Similarly to Amazon Biobank, they argue that permissioned blockchains (e.g., Hyperledger Fabric) better meet the complex requirements for genetic data storage, such as user identification. It supports the purchase of genomic data access both in fiat and cryptocurrency transfers, using third-party APIs such as Stellar and Strip. In addition, Genesy aims to eventually aggregate more organizations in their consortium, strengthening their open governance model and encouraging collaboration and fairness.

It is worth noting that Nebula Genomic, Genesy, and other platforms prioritize human genetics, and thus, they do not focus on biodiversity as an asset. Consequently, they provide limited support for intellectual property protection or benefit-sharing. Additionally, to safeguard user privacy, these platforms often restrict the identification of data owners to the company or Federation only. While this is appropriate in the context of human genetics, it is not desirable in the case of the Amazon Biobank. Also, the anonymization of data owners hinders the preservation of their intellectual property rights and restricts their fair compensation.

In the context of non-human genetic data, in 2021, the United Nations Development Programme (UNDP) conducted a blockchain-based project to improve genomic resource traceability and benefit-sharing (UNDP, 2021). With the major goal of implementing the Nagoya Protocol (BUCK; HAMILTON, 2011), the Global ABS Tracker project is presently in the early stages, with a launched pilot prototype. The project, nonetheless, tries to handle all kinds of natural products, like plants or natural substances, not focusing on genetic data only. Hence the system does not support collaborative and private storage of genomic data, nor DNA analysis, validation, and sequence search. In addition, one of the challenges of the project is that it requires global coordination between countries, something that is still a work in progress.

7 Conclusion

In this work, we present the Amazon Biobank, a community-based genetic database that implements monetary incentives for users who collaborate with data, knowledge, and computational resources. The resulting system provides strong traceability and auditability features, making it easier to link biotechnology assets to registered data and to verify compliance with data usage and benefit-sharing agreements. In addition, by leveraging collaborative technologies like BitTorrent and blockchain, the proposed architecture becomes highly scalable and less dependent on the trust deposited in any particular system player.

Our system serves as an alternative to several existing databases that register biodiversity genetic data, such as NCBI and EBI. Despite the relevance of those repositories, they lack adequate sharing of economic benefits resulting from exploring genomes. In our solution, people with easy access to high-biodiversity areas, such as local community members, are encouraged to insert genetic data. This will increase the variability of DNA data cataloged, especially in challenging and extensive areas such as the Amazon Rainforest.

Therefore, we expect that the Amazon Biobank can be an important stepping-stone to unlocking the huge potential of bioeconomy in rich ecosystems such as the Amazon Rainforest. In particular, it should foster innovations via biomimetic engineering, synthetic biology, or new materials development, besides bringing social impact to the local communities via sustainable economic development.

7.1 Publications

Resulting of the research carried out during this work, we produced the following publications:

- Journal Article: KIMURA, L. T. et al. Amazon biobank: a collaborative genetic database for bioeconomy development. *Functional & Integrative Genomics*, v. 23, n. 2, p. 101, Mar 2023. ISSN 1438-7948. Available at: <https://doi.org/10.1007/s10142-023-01015-1>.

- Conference Paper: KIMURA, L. et al. Amazon Biobank: sustainable development built upon rainforest's biodiversity. In: *Planetary Health Annual Meeting and Festival*. Brasil: PHAM, 2021. Available at: <https://bit.ly/3eyLht9>.

- Conference Paper: KIMURA, L. et al. Amazon biobank - a community-based genetic database. In: *Proc. of the XXI Brazilian Symposium on Information and*

Computational Systems Security (SBSeg). Porto Alegre/RS, Brazil: SBC, 2021. p. 74–81.

We also produced the following work during the master's degree.

- Conference Paper: KIMURA, L. et al. Logs transparentes: transparência e auditabilidade usando estruturas de dados verificáveis. In: *Anais Estendidos do XXII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*. Porto Alegre, RS, Brasil: SBC, 2022. p. 87–94. Available at: <https://sol.sbc.org.br/index.php/sbseg_estendido/article/view/21696>.

7.2 Future Work

The next steps for Amazon Biobank include deploying the prototype in the Amazon region, first still as a small-scale demonstration. It would include presenting the prototype to target populations (e.g., traditional community members) and populating it with real genetic data to collect feedback. Thus, we intend to make the system as user-friendly as possible, avoiding Collectors having to rely on others to insert the data. This deployment most probably would include collaboration with other universities or non-profit organizations (such as Amazon 4.0 initiative[1]).

After those tests, we plan to execute a larger-scale assessment. This would include establishing an intercommunication channel with existing systems that mediate and lay the legal grounds for the usage of genetic data in the Amazon Forest, like the Brazilian National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SISGEN, 2021). Other non-technological aspects also must be emphasized, such as regulations on biodiversity and financial assets. We expect that this deployment will result in a more mature specification that can be used to implement the definitive version of Amazon Biobank.

From a technical standpoint, some functionalities within the Amazon Biobank could benefit from further elaboration and implementation. For example, improving the sequence search operation could involve introducing a distributed approach, such as using a Distributed Hash Table (DHT), or incorporating privacy-preserving mechanisms like searchable encryption. Additionally, another challenge is to establish a robust reputation system that can address multiple malicious participants, including Validators and Curators. One potential solution is to explore blockchain-based reputation mechanisms to enhance system transparency and accountability.

[1] <https://amazonia4.org/>

Bibliography

ALGHAZWI, M. et al. Blockchain for genomics: A systematic literature review. *Distrib. Ledger Technol.*, Association for Computing Machinery, New York, NY, USA, v. 1, n. 2, dec 2022. ISSN 2769-6472. Available at: <https://doi.org/10.1145/3563044>. Cited on page 14.

ANDROULAKI, E. et al. Hyperledger fabric: A distributed operating system for permissioned blockchains. In: *Proceedings of the Thirteenth EuroSys Conference*. New York, NY, USA: Association for Computing Machinery, 2018. (EuroSys '18). ISBN 9781450355841. Available at: <https://doi.org/10.1145/3190508.3190538>. Cited on page 20.

ANTWI, M. et al. The case of hyperledger fabric as a blockchain solution for healthcare applications. *Blockchain: Research and Applications*, v. 2, n. 1, p. 100012, 2021. ISSN 2096-7209. Available at: <https://www.sciencedirect.com/science/article/pii/S2096720921000075>. Cited on page 19.

BACH, L. M.; MIHALJEVIC, B.; ZAGAR, M. Comparative analysis of blockchain consensus algorithms. *41st Int. Convention on Information and Communication Technology, Electronics and Microelectronics*, p. 1545–1550, 2018. Cited on page 25.

BATTISTI, J. H. F. et al. Analysis of an ethereum private blockchain network hosted by virtual machines against internal dos attacks. In: BAROLLI, L.; HUSSAIN, F.; ENOKIDO, T. (Ed.). *Advanced Information Networking and Applications*. Cham: Springer International Publishing, 2022. p. 479–490. ISBN 978-3-030-99584-3. Cited on page 50.

BEECH, E. et al. GlobalTreeSearch: The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry*, Taylor & Francis, v. 36, n. 5, p. 454–489, 2017. Cited on page 12.

BENET, J. *IPFS - Content Addressed, Versioned, P2P File System*. 2014. Cited on page 21.

BEYENE, M. et al. A scoping review of distributed ledger technology in genomics: thematic analysis and directions for future research. *Journal of the American Medical Informatics Association*, v. 29, n. 8, p. 1433–1444, 05 2022. ISSN 1527-974X. Available at: <https://doi.org/10.1093/jamia/ocac077>. Cited on page 14.

BitTorrent Foundation. *BitTorrent Speed*. 2019. <https://www.bittorrent.com/token/bittorrent-speed/>. (online). Cited on page 21.

BRONDÍZIO, E. S. The amazonian caboclo and the açaí palm: forest farmers in the global market. *Advances in Economic Botany*, JSTOR, v. 16, p. iii–403, 2008. Cited on page 13.

BUCK, M.; HAMILTON, C. The Nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. *Review of ECIEL*, Wiley Online Library, v. 20, n. 1, p. 47–61, 2011. Cited on page 52.

BUTERIN, V. et al. A next-generation smart contract and decentralized application platform. *white paper*, v. 3, n. 37, p. 2–1, 2014. Cited on page 17.

CARLINI, R. et al. Genesy: a blockchain-based platform for DNA sequencing. *DLT@ ITASEC*, v. 2019, p. 68–72, 2019. Cited 2 times on pages 19 and 52.

CHASE, B.; MACBROUGH, E. Analysis of the XRP ledger consensus protocol. *arXiv preprint arXiv:1802.07242*, 2018. Cited on page 18.

COHEN, B. Incentives build robustness in BitTorrent. In: *Workshop on Economics of Peer-to-Peer systems*. 2003. v. 6, p. 68–72. Cited 2 times on pages 21 and 29.

DABHOLKAR, A.; SARASWAT, V. Ripping the fabric: Attacks and mitigations on hyperledger fabric. In: SPRINGER. *International Conference on Applications and Techniques in Information Security*. 2019. p. 300–311. Cited on page 48.

DHILLON, V.; METCALF, D.; HOOPER, M. The Hyperledger project. In: *Blockchain enabled applications*. : Springer, 2017. p. 139–149. Cited on page 40.

DOMINGUES, A. F. et al. Pigmentos antociânicos do açaí (euterpe oleracea mart.) como evidenciadores de biofilme dental. In: PESSOA, JDC; TEIXEIRA, GH de A.(Ed.). Tecnologias para inovação nas . . . , 2012. Cited on page 13.

ENCRYPGEN. The clinical and investment potential in the gene-chain project the unprecedented growth of genomic data. 2017. Available at: <http://icotimeline.com/wp-content/uploads/2017/07/Gene-Chain-Whitepaper.pdf>. Cited on page 51.

EYAL, I.; SIRER, E. Majority is not enough: Bitcoin mining is vulnerable. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 61, n. 7, p. 95—-102, 2018. ISSN 0001-0782. Available at: <https://doi.org/10.1145/3212998>. Cited on page 48.

FU, W.; WEI, X.; TONG, S. An improved blockchain consensus algorithm based on raft. *Arabian Journal for Science and Engineering*, v. 46, n. 9, p. 8137–8149, Sep 2021. ISSN 2191-4281. Available at: <https://doi.org/10.1007/s13369-021-05427-8>. Cited on page 21.

FUKUDA, A. et al. DDBJ update: streamlining submission and access of human data. *Nucleic Acids Research*, Oxford University Press (OUP), v. 49, n. D1, p. D71–D75, Nov. 2021. Available at: <https://doi.org/10.1093/nar/gkaa982>. Cited on page 13.

GLOWKA, L. et al. A guide to the convention on biological diversity. IUCN: International Union for Conservation of Nature, 1994. Cited on page 13.

GOLDSTEIN, P. Z.; DESALLE, R. Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays*, Wiley Online Library, v. 33, n. 2, p. 135–147, 2011. Cited on page 27.

GRISHIN, D. et al. *Blockchain-enabled genomic data sharing and analysis platform*. 2018. Cited on page 52.

GUGGENBERGER, T. et al. An in-depth investigation of the performance characteristics of hyperledger fabric. *Computers & Industrial Engineering*, v. 173, p. 108716, 2022. ISSN 0360-8352. Available at: <https://www.sciencedirect.com/science/article/pii/S0360835222007045>. Cited 2 times on pages 33 and 43.

HARRISON, P. W. et al. The European nucleotide archive in 2020. *Nucleic Acids Research*, Oxford University Press (OUP), v. 49, n. D1, p. D82–D85, Nov. 2021. Available at: <https://doi.org/10.1093/nar/gkaa1028>. Cited on page 13.

HAZEL, G.; ARVID, N. *Extension for Peers to Send Metadata Files.* 2008. Available at: <https://www.bittorrent.org/beps/bep_0009.html>. Cited on page 21.

HOORN, C. et al. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*, v. 330, n. 6006, p. 927–931, 2010. Available at: <https://www.science.org/doi/abs/10.1126/science.1194585>. Cited on page 12.

HUANG, D.; MA, X.; ZHANG, S. Performance analysis of the raft consensus algorithm for private blockchains. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, v. 50, n. 1, p. 172–181, 2020. Cited on page 21.

HYPERLEDGER Caliper repository. 2022. Available at: <https://github.com/hyperledger/caliper>. Cited on page 41.

HYPERLEDGER Explorer repository. 2022. Available at: <https://github.com/hyperledger-labs/blockchain-explorer>. Cited on page 36.

ITO, K.; O'DAIR, M. A critical examination of the application of blockchain technology to intellectual property management. In: *Business transformation through blockchain.* : Springer, 2019. p. 317–335. Cited on page 49.

JAIN, M. et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, Springer, v. 17, n. 1, p. 1–11, 2016. Cited on page 36.

KIMURA, L. et al. Amazon biobank - a community-based genetic database. In: *Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg).* Porto Alegre/RS, Brazil: SBC, 2021. p. 74–81. Cited on page 54.

KIMURA, L. et al. Amazon Biobank: sustainable development built upon rainforest's biodiversity. In: *Planetary Health Annual Meeting and Festival.* Brasil: PHAM, 2021. Available at: <https://bit.ly/3eyLht9>. Cited on page 53.

KIMURA, L. et al. Logs transparentes: transparência e auditabilidade usando estruturas de dados verificáveis. In: *Anais Estendidos do XXII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais.* Porto Alegre, RS, Brasil: SBC, 2022. p. 87–94. Available at: <https://sol.sbc.org.br/index.php/sbseg_estendido/article/view/21696>. Cited on page 54.

KIMURA, L. T. et al. Amazon biobank: a collaborative genetic database for bioeconomy development. *Functional & Integrative Genomics*, v. 23, n. 2, p. 101, Mar 2023. ISSN 1438-7948. Available at: <https://doi.org/10.1007/s10142-023-01015-1>. Cited on page 53.

KULEMIN, N.; POPOV, S.; GORBACHEV, A. The zenome project: Whitepaper blockchain-based genomic ecosystem. *Zenome. io*, p. A, 2017. Cited on page 51.

LAURIE, B. Certificate transparency. *Communications of the ACM*, ACM New York, NY, USA, v. 57, n. 10, p. 40–46, 2014. Cited on page 18.

LEVIS, C. et al. Persistent effects of pre-columbian plant domestication on amazonian forest composition. *Science*, v. 355, n. 6328, p. 925–931, 2017. Available at: <https://www.science.org/doi/abs/10.1126/science.aal0157>. Cited on page 12.

LEVIS, C. et al. How people domesticated amazonian forests. *Frontiers in Ecology and Evolution*, v. 5, p. 171, 2018. ISSN 2296-701X. Available at: <https://www.frontiersin.org/article/10.3389/fevo.2017.00171>. Cited on page 12.

LI, D.; WONG, W. E.; GUO, J. A survey on blockchain for enterprise using hyperledger fabric and composer. In: *2019 6th International Conference on Dependable Systems and Their Applications (DSA)*. 2020. p. 71–80. Cited 2 times on pages 18 and 19.

LI, F.-W. Decolonizing botanical genomics. *Nature Plants*, v. 7, n. 12, p. 1542–1543, Dec 2021. ISSN 2055-0278. Available at: <https://doi.org/10.1038/s41477-021-01041-6>. Cited on page 13.

LOEWENSTERN, A.; ARVID, N. *DHT Protocol*. 2008. Available at: <http://www.bittorrent.org/beps/bep_0005.html>. Cited on page 21.

MAYMOUNKOV, P.; MAZIÈRES, D. Kademlia: A peer-to-peer information system based on the XOR metric. In: *Peer-to-Peer Systems*. Berlin, Heidelberg: Springer, 2002. p. 53–65. ISBN 978-3-540-45748-0. Cited on page 21.

MICROSOFT. *What is Delivery Optimization*. 2022. Accessed in 10-02-2023. Available at: <https://learn.microsoft.com/en-us/windows/deployment/do/waas-delivery-optimization>. Cited on page 21.

MIT. *A glossary of blockchain jargon*. 2018. MIT Technology Review. Available at: <https://www.technologyreview.com/2018/04/23/143486/a-glossary-of-blockchain-jargon/>. Cited on page 17.

MYERS, N. et al. Biodiversity hotspots for conservation priorities. *Nature*, Nature Publishing Group, v. 403, n. 6772, p. 853–858, 2000. Cited on page 12.

NA, D. Dna steganography: hiding undetectable secret messages within the single nucleotide polymorphisms of a genome and detecting mutation-induced errors. *Microbial Cell Factories*, v. 19, n. 1, p. 128, Jun 2020. ISSN 1475-2859. Available at: <https://doi.org/10.1186/s12934-020-01387-0>. Cited on page 50.

NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, p. 21260, 2008. Cited on page 17.

NASI, R.; BILLAND, A.; van Vliet, N. Managing for timber and biodiversity in the congo basin. *Forest Ecology and Management*, v. 268, p. 103–111, 2012. ISSN 0378-1127. Multiple Use of Tropical Forests: From Concept to Reality. Available at: <https://www.sciencedirect.com/science/article/pii/S0378112711002209>. Cited on page 12.

NOBRE, C. A. et al. Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proc. of the National Academy of Sciences*, National Academy of Sciences, v. 113, n. 39, p. 10759–10768, 2016. ISSN 0027-8424. Available at: <https://www.pnas.org/content/113/39/10759>. Cited on page 12.

NOBRE, I.; NOBRE, C. A. The Amazonia third way initiative: the role of technology to unveil the potential of a novel tropical biodiversity-based economy. *Land use. Assessing the Past, Envisioning the Future*, 2019. Cited 3 times on pages 12, 13, and 15.

ONGARO, D.; OUSTERHOUT, J. In search of an understandable consensus algorithm. In: *Proc. of the 2014 USENIX Conference on USENIX Annual Technical Conference*. USA: USENIX Association, 2014. p. 305–320. ISBN 9781931971102. Cited on page 18.

ONGARO, D.; OUSTERHOUT, J. In search of an understandable consensus algorithm. In: *USENIX Annual Technical Conference*. 2014. p. 305–319. Cited on page 20.

OZERCAN, H. I. et al. Realizing the potential of blockchain technologies in genomics. *Genome research*, Cold Spring Harbor Lab, v. 28, n. 9, p. 1255–1263, 2018. Cited 2 times on pages 14 and 18.

PORTINHO, J. A.; ZIMMERMANN, L. M.; BRUCK, M. R. Efeitos benéficos do açaí. *International journal of nutrology*, Thieme Revinter Publicações Ltda, v. 5, n. 01, p. 015–020, 2012. Cited on page 13.

RECH, E. Genomics and synthetic biology as a viable option to intensify sustainable use of biodiversity. *Nature Precedings*, Mar 2011. ISSN 1756-0357. Available at: <https://doi.org/10.1038/npre.2011.5759.1>. Cited on page 13.

RINTELEN, K. V.; ARIDA, E.; HÄUSER, C. A review of biodiversity-related issues and challenges in megadiverse indonesia and other southeast asian countries. *Research Ideas and Outcomes*, Pensoft Publishers, v. 3, p. e20860, 2017. Cited on page 12.

RIVEST, R. L.; SHAMIR, A. Payword and micromint: Two simple micropayment schemes. In: *Security Protocols*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. p. 69–87. ISBN 978-3-540-68047-5. Cited on page 21.

ROBINSON, P.; BRAINARD, J. Anonymous state pinning for private blockchains. In: IEEE. *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. 2019. p. 827–834. Cited on page 48.

ROSE, S. et al. *Zero trust architecture*. 2020. Cited on page 15.

SAYERS, E. W. et al. GenBank. *Nucleic Acids Research*, Oxford University Press (OUP), v. 49, n. D1, p. D92–D96, Nov. 2020. Available at: <https://doi.org/10.1093/nar/gkaa1023>. Cited on page 13.

SEPPEY, M.; MANNI, M.; ZDOBNOV, E. M. BUSCO: Assessing genome assembly and annotation completeness. In: _____. *Gene Prediction: Methods and Protocols*. New York, NY: Springer, 2019. p. 227–245. ISBN 978-1-4939-9173-0. Available at: <https://doi.org/10.1007/978-1-4939-9173-0_14>. Cited 3 times on pages 27, 30, and 49.

SHIMOSAWA, T.; SATO, T.; OSHIMA, S. Bcverifier: A tool to verify hyperledger fabric ledgers. In: *2020 IEEE International Conference on Blockchain (Blockchain)*. 2020. p. 291–299. Cited on page 48.

SHIRAISHI, F. et al. Torrente, a micropayment based Bittorrent extension to mitigate free riding. In: *Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg)*. Porto Alegre/RS, Brazil: SBC, 2021. p. 82–89. Cited 4 times on pages 16, 21, 29, and 39.

SHRIVAS, M. The disruptive blockchain: Types, platforms and applications. *Texila Int. Journal of Academic Research*, p. 17–39, 04 2019. Cited on page 17.

SILVA, M. F. d. O.; PEREIRA, F. d. S.; MARTINS, J. V. B. A bioeconomia brasileira em números. *BNDS Setorial*, Rio de Janeiro, p. [277]–311, mar 2018. Cited on page 13.

SISGEN. *SisGen - Sistema Nacional de Gestão de Patrimônio Genético e do Conhecimento Tradicional Associado*. 2021. <https://www.mma.gov.br/patrimonio-genetico/conselho-de-gestao-do-patrimonio-genetico/sis-gen>. [Online; accessed 23-February-2021]. Cited on page 54.

STAFFAS, L.; GUSTAVSSON, M.; MCCORMICK, K. Strategies and policies for the bioeconomy and bio-based economy: An analysis of official national approaches. *Sustainability*, v. 5, n. 6, p. 2751–2769, 2013. ISSN 2071-1050. Available at: <https://www.mdpi.com/2071-1050/5/6/2751>. Cited on page 12.

STEEGE, H. ter et al. Hyperdominance in the Amazonian tree flora. *Science*, v. 342, n. 6156, p. 1243092, 2013. Available at: <https://www.science.org/doi/abs/10.1126/science.1243092>. Cited on page 12.

SWAN, M. *Blockchain: Blueprint for a new economy.* : O'Reilly Media, Inc., 2015. Cited on page 17.

THIEBES, S. et al. Beyond data markets: Opportunities and challenges for distributed ledger technology in genomics. *Proc. of the 53rd Hawaii Int. Conf. on System Sciences*, v. 3, p. 3275–3284, 2020. Cited on page 14.

UNDP. *A pilot to improve genetic resources traceability through blockchain technology launched by the UNDP GEF Global ABS Project*. 2021. <https://bit.ly/3hnMqEh>. Acessed on 29-06-2021. Cited on page 52.

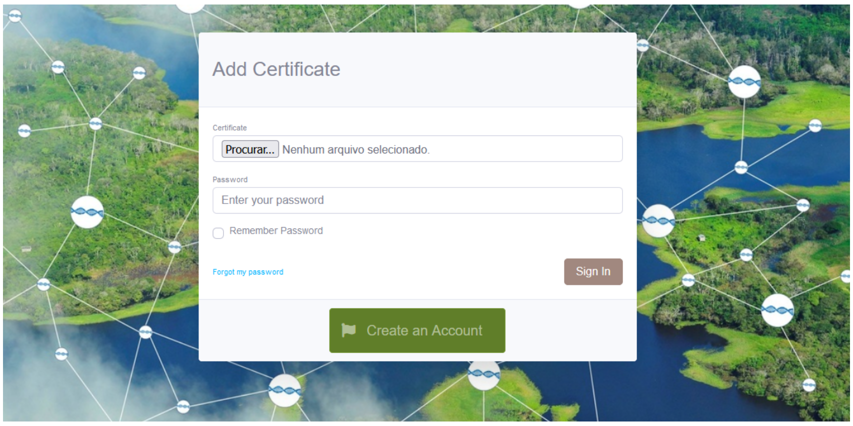
WIPO. *Traditional Knowledge*. 2022. World Intellectual Property Organization - <https://www.wipo.int/tk/en/tk/> (online). Cited on page 28.

APPENDIX A – Tutorials for Biobank Interface

A.1 Login

1. Open the biobank-app, and go to <http://127.0.0.1:3000> in your browser
2. Insert the certificate and the password
3. Click in Sign in

Figure 18 – Amazon Biobank - login page

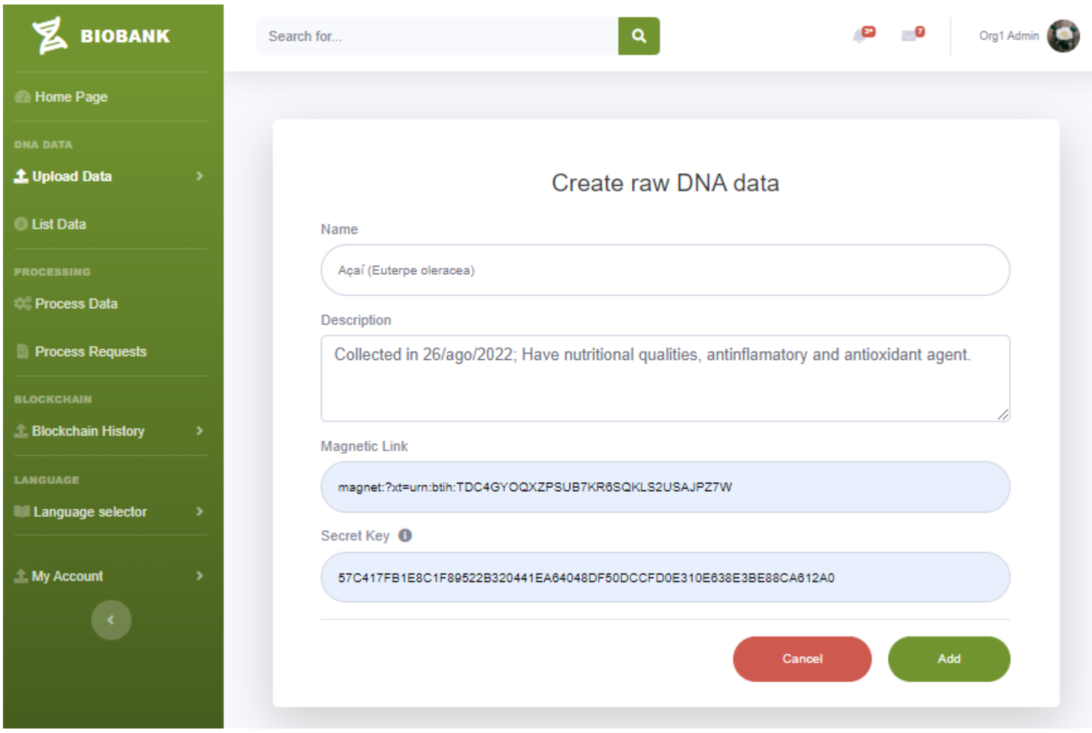


Source: Author

A.2 Insert DNA data

1. On Biobank-app homepage, click in "insert DNA", and then "Raw Data"
2. Fill in the "name" and the "description" fields with information about the DNA data
   2.1 In our case, we inserted "Açaí (Euterpe oleracea)" as name, and "Collected in 26/ago/2022. Have nutritional qualities, antinflamatory and antioxidant agent" as a description
3. Get the magnet link and the decryption key in Torrente, and copy+paste here
4. Click on "add" to save the data on the blockchain

Figure 19 – Inserting raw DNA data



Create raw DNA data

Name: Açaí (Euterpe oleracea)

Description: Collected in 26/ago/2022; Have nutritional qualities, antinflamatory and antioxidant agent.

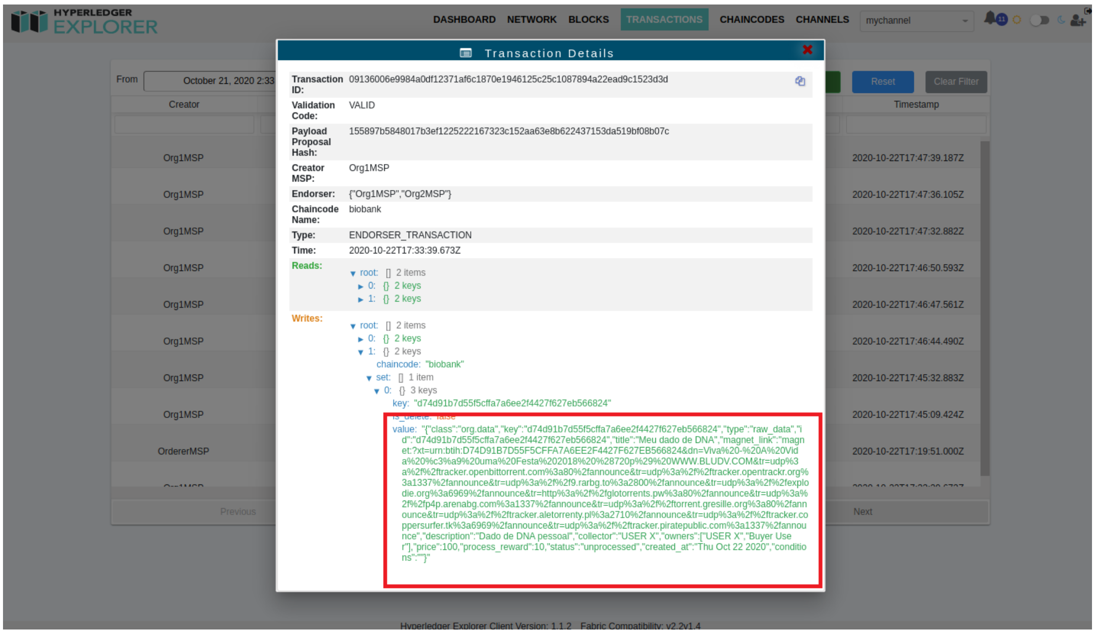
Magnetic Link: magnet:?xt=urn:btih:TDC4GYOQXZPSUB7KR6SQKLS2USAJPZ7W

Secret Key: 57C417FB1E8C1F89522B320441EA64048DF50DCCFD0E310E638E3BE88CA612A0

Source: Author

Now, this genetic data and the associated metadata are registered on the blockchain. You can see the details by visiting http://amazonbiobank.duckdns.org:8080 -> transactions and looking at the details.

Figure 20 – Transaction details in Hyperledger Explorer. The data written on the blockchain is marked in red.



Transaction Details

Transaction ID:	09136006e9984a0df12371af6c1870e1946125c25c1087894a22ead9c1523d3d
Validation Code:	VALID
Payload Proposal Hash:	155897b5848017b3ef1225222167323c152aa63e8b622437153da519bf08b07c
Creator MSP:	Org1MSP
Endorser:	{"Org1MSP","Org2MSP"}
Chaincode Name:	biobank
Type:	ENDORSER_TRANSACTION
Time:	2020-10-22T17:33:39.673Z

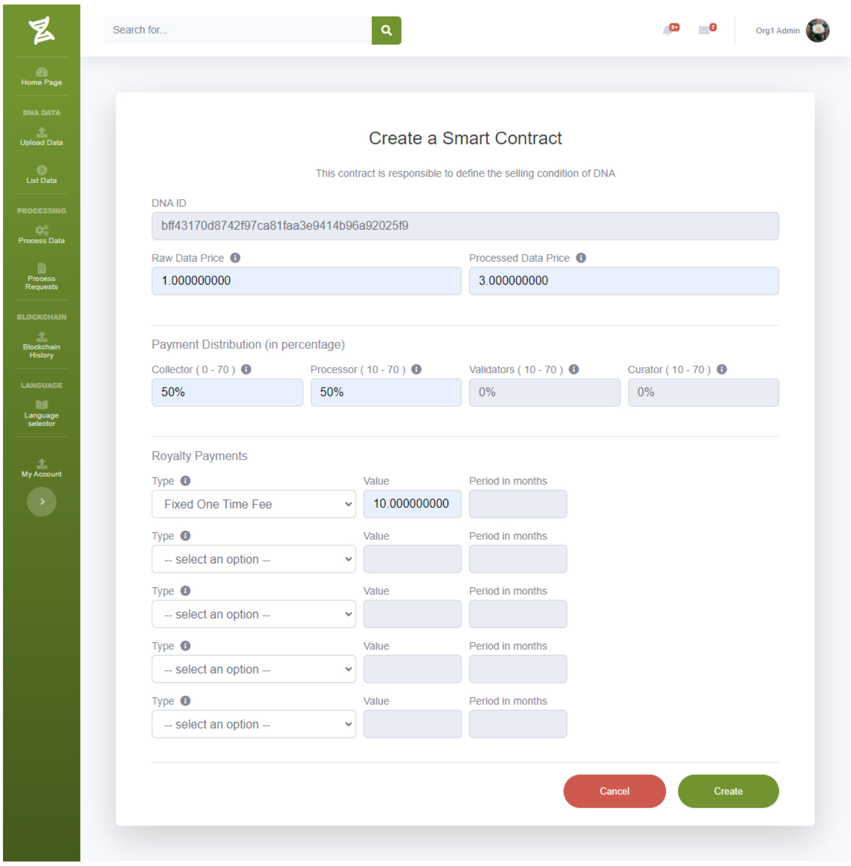
```
Reads:
▼ root: [] 2 items
  ▶ 0: {} 2 keys
  ▶ 1: {} 2 keys
Writes:
▼ root: [] 2 items
  ▶ 0: {} 2 keys
  ▼ 1: {} 2 keys
      chaincode: "biobank"
    ▼ set: [] 1 item
      ▼ 0: {} 3 keys
          key: "d74d91b7d55f5cffa7a6ee2f4427f627eb566824"
          is_delete: false
          value: "{"class":"org.data","key":"d74d91b7d55f5cffa7a6ee2f4427f627eb566824","type":"raw_data","id":"d74d91b7d55f5cffa7a6ee2f4427f627eb566824","title":"Meu dado de DNA","magnet_link":"magnet:?xt=urn:btih:D74D91B7D55F5CFFA7A6EE2F4427F627EB566824&dn=Viva%20-%20A%20Vida%20%c3%a9%20uma%20Festa%202018%20%28720p%29%20WWW.BLUDV.COM&tr=udp%3a%2f%2ftracker.openbittorrent.com%3a80%2fannounce&tr=udp%3a%2f%2ftracker.opentrackr.org%3a1337%2fannounce&tr=udp%3a%2f%2f9.rarbg.to%3a2800%2fannounce&tr=udp%3a%2f%2fexplodie.org%3a6969%2fannounce&tr=http%3a%2f%2fglotorrents.pw%3a80%2fannounce&tr=udp%3a%2f%2fp4p.arenabg.com%3a1337%2fannounce&tr=udp%3a%2f%2ftorrent.gresille.org%3a80%2fannounce&tr=udp%3a%2f%2ftracker.aletorrenty.pl%3a2710%2fannounce&tr=udp%3a%2f%2ftracker.coppersurfer.tk%3a6969%2fannounce&tr=udp%3a%2f%2ftracker.piratepublic.com%3a1337%2fannounce","description":"Dado de DNA pessoal","collector":"USER X","owners":["USER X","Buyer User"],"price":100,"process_reward":10,"status":"unprocessed","created_at":"Thu Oct 22 2020","conditions":""}"
```

Source: Author

A.3 Insert smart contract parameters

1. After inserting raw DNA data, click on "create smart contract"
2. Insert the selling conditions of DNA
   a) Those include "raw data price", "processed data price", and the payment distribution
   b) You can also include the royalty payment method that you accept. This will be presented to Buyers when buying access to your data.
3. Click in "create" to save your parameters

Figure 21 – Creating a smart contract for raw DNA data. Inserted parameters are raw data price (1 biocoins); processed data price (3 biocoins); payment distribution (50% Collectors; 50% Processors); and royalty payment (fixed one-time fee of 10 biocoins).



Create a Smart Contract

This contract is responsible to define the selling condition of DNA

DNA ID: bff43170d8742f97ca81faa3e9414b96a92025f9

Raw Data Price: 1.000000000

Processed Data Price: 3.000000000

Payment Distribution (in percentage)

Collector (0 - 70)	Processor (10 - 70)	Validators (10 - 70)	Curator (10 - 70)
50%	50%	0%	0%

Royalty Payments

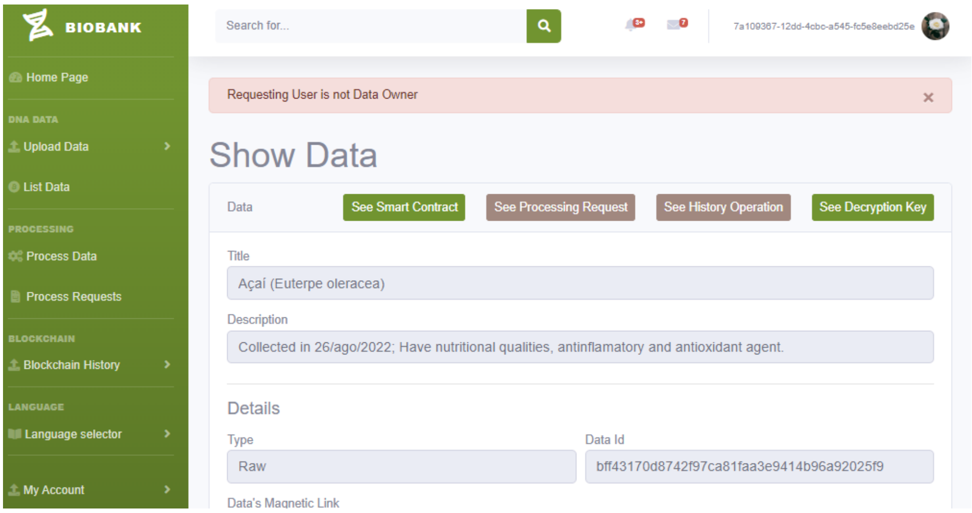
Type	Value	Period in months
Fixed One Time Fee	10.000000000	
-- select an option --		
-- select an option --		
-- select an option --		
-- select an option --		

Source: Author

A.4 Buy raw DNA data

1. Login in Biobank-app (possibly using another account)
2. Before buying the data, we will test that you cannot access the data encryption key. Go to "list data", select the data you want to buy, and click on "see decryption key". You must get an error message. (Figure 22)

Figure 22 – Denied attempt to access the data decryption key



Requesting User is not Data Owner

Show Data

Data

Title: Açaí (Euterpe oleracea)

Description: Collected in 26/ago/2022; Have nutritional qualities, antinflamatory and antioxidant agent.

Details

Type: Raw

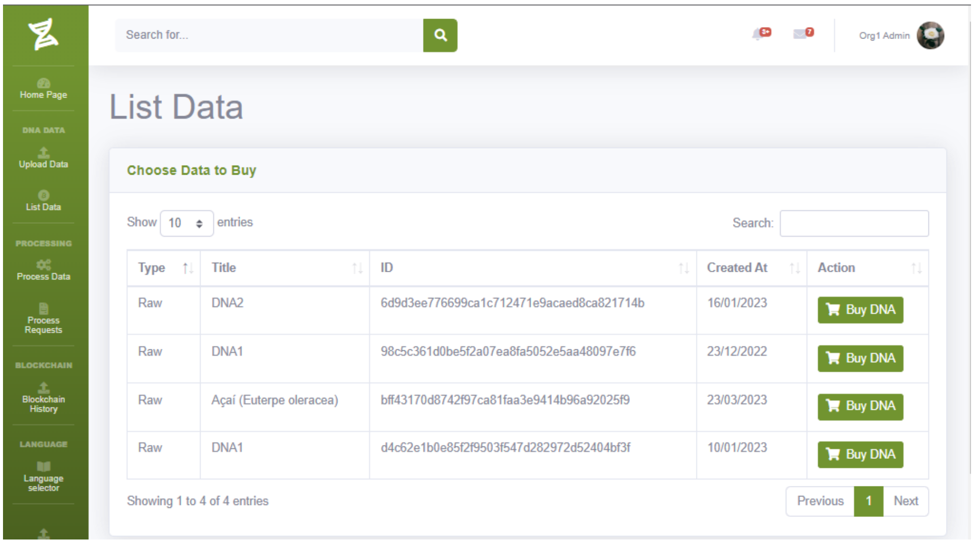
Data Id: bff43170d8742f97ca81faa3e9414b96a92025f9

Data's Magnetic Link

Source: Author

3. Now, let's buy the DNA data. On the home page, click on "buy data". Then, select the DNA data you want to buy (Figure 23)

Figure 23 – List of DNA available for purchase



List Data

Choose Data to Buy

Show 10 entries

Search:

Type	Title	ID	Created At	Action
Raw	DNA2	6d9d3ee776699ca1c712471e9acaed8ca821714b	16/01/2023	Buy DNA
Raw	DNA1	98c5c361d0be5f2a07ea8fa5052e5aa48097e7f6	23/12/2022	Buy DNA
Raw	Açaí (Euterpe oleracea)	bff43170d8742f97ca81faa3e9414b96a92025f9	23/03/2023	Buy DNA
Raw	DNA1	d4c62e1b0e85f2f9503f547d282972d52404bf3f	10/01/2023	Buy DNA

Showing 1 to 4 of 4 entries

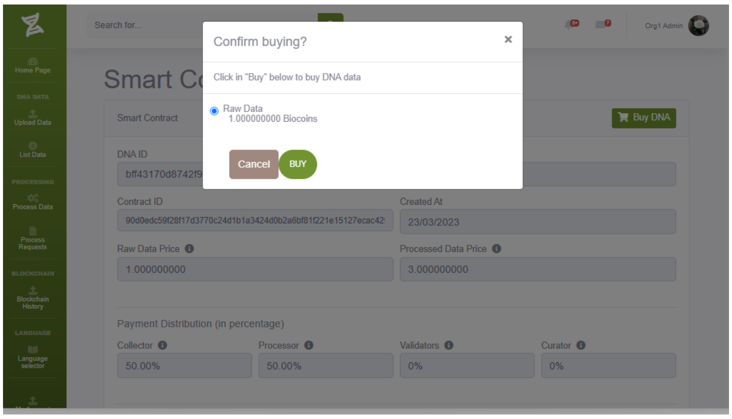
Source: Author

4. After seeing the DNA data detail, click on "see Smart Contract"

5. After seeing the buying conditions (including the raw data price), click on "Buy DNA"

6. Click "buy" on the confirmation screen. (Figure 24)

Figure 24 – Buying raw DNA data - confirmation screen



Confirm buying?

Click in "Buy" below to buy DNA data

● Raw Data
1.000000000 Biocoins

Cancel BUY

Smart Contract

DNA ID: bff43170d8742f9

Contract ID: 90d0edc59f28f17d3770c24d1b1a3424d0b2a6bf81f221e15127ecac42

Created At: 23/03/2023

Raw Data Price: 1.000000000

Processed Data Price: 3.000000000

Payment Distribution (in percentage)

Collector: 50.00%

Processor: 50.00%

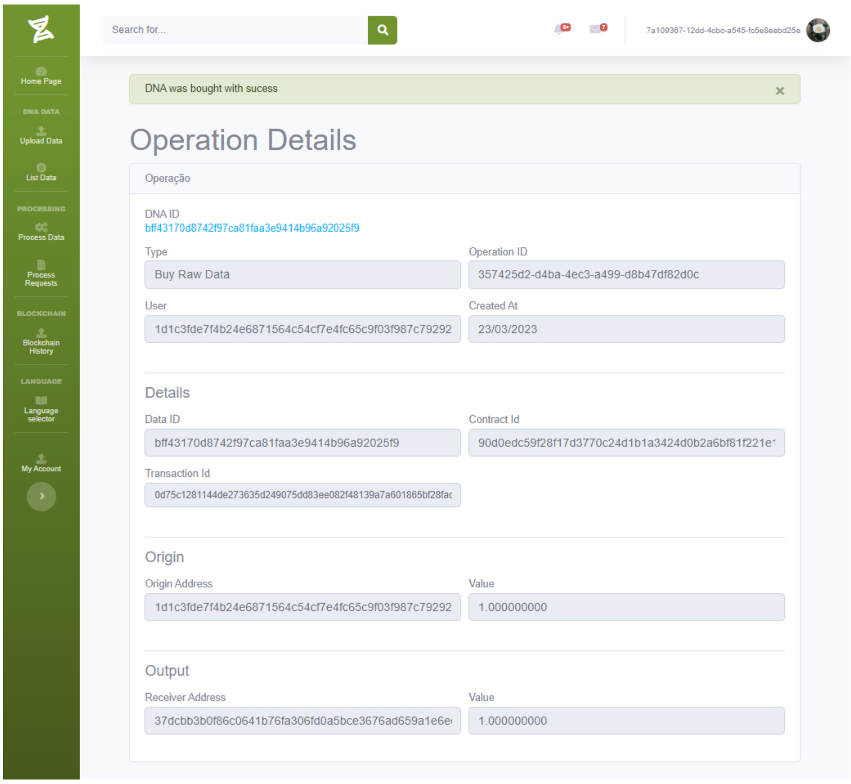
Validators: 0%

Curator: 0%

Source: Author

7. After the payment, you will see the "operation receipt", which contains your userID, the value in biocoins that were transferred, and the destination address (Figure 25)

Figure 25 – Details about the buying operation. The value that was transferred, destination userID, and DNA dataID



DNA was bought with sucess

Operation Details

Operação

DNA ID: bff43170d8742f97ca81faa3e9414b96a92025f9

Type: Buy Raw Data

Operation ID: 357425d2-d4ba-4ec3-a499-d8b47df82d0c

User: 1d1c3fde7f4b24e6871564c54cf7e4fc65c9f03f987c79292

Created At: 23/03/2023

Details

Data ID: bff43170d8742f97ca81faa3e9414b96a92025f9

Contract Id: 90d0edc59f28f17d3770c24d1b1a3424d0b2a6bf81f221e1

Transaction Id: 0d75c1281144de273635d249075dd83ee082f48139a7a601865bf28fac

Origin

Origin Address: 1d1c3fde7f4b24e6871564c54cf7e4fc65c9f03f987c79292

Value: 1.000000000

Output

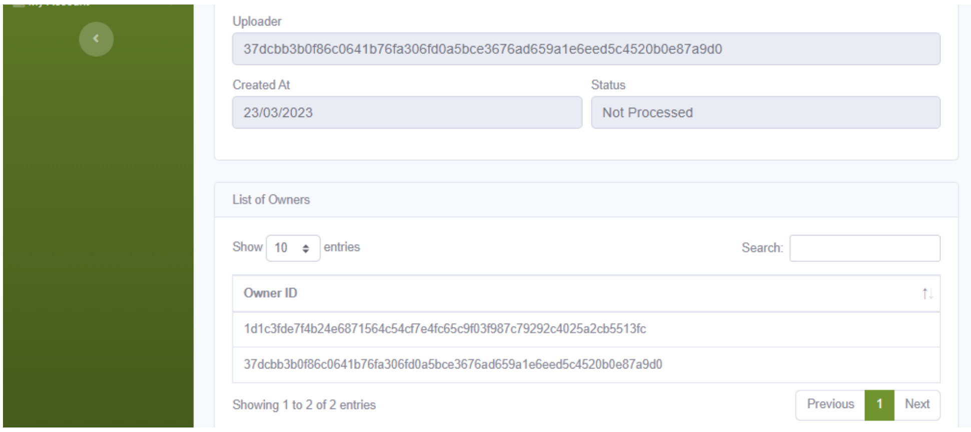
Receiver Address: 37dcbb3b0f86c0641b76fa306fd0a5bce3676ad659a1e6e

Value: 1.000000000

Source: Author

8. On the operation receipt, click on the link with DNA ID
9. In the DNA data detail page, note that your userID has been included on the "owner list" (Figure 26)

Figure 26 – Updated list of owners. Note that your userID has been included



Uploader
37dcbb3b0f86c0641b76fa306fd0a5bce3676ad659a1e6eed5c4520b0e87a9d0

Created At
23/03/2023

Status
Not Processed

List of Owners

Show 10 entries Search:

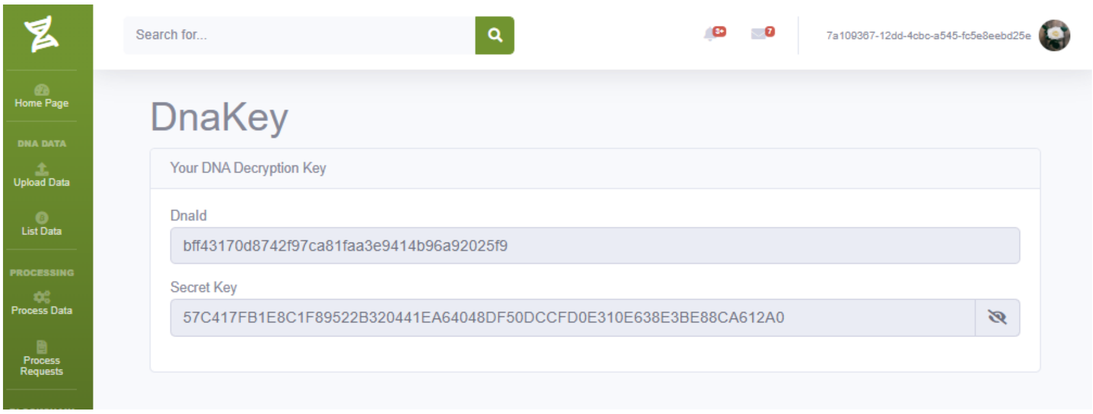
Owner ID
1d1c3fde7f4b24e6871564c54cf7e4fc65c9f03f987c79292c4025a2cb5513fc
37dcbb3b0f86c0641b76fa306fd0a5bce3676ad659a1e6eed5c4520b0e87a9d0

Showing 1 to 2 of 2 entries

Source: Author

10. Click on "See decryption key". Since you are now one of the DNA data owners, you have received access to it. (Figure 27)

Figure 27 – DNA data decryption key



DnaKey

Your DNA Decryption Key

DnaId
bff43170d8742f97ca81faa3e9414b96a92025f9

Secret Key
57C417FB1E8C1F89522B320441EA64048DF50DCCFD0E310E638E3BE88CA612A0

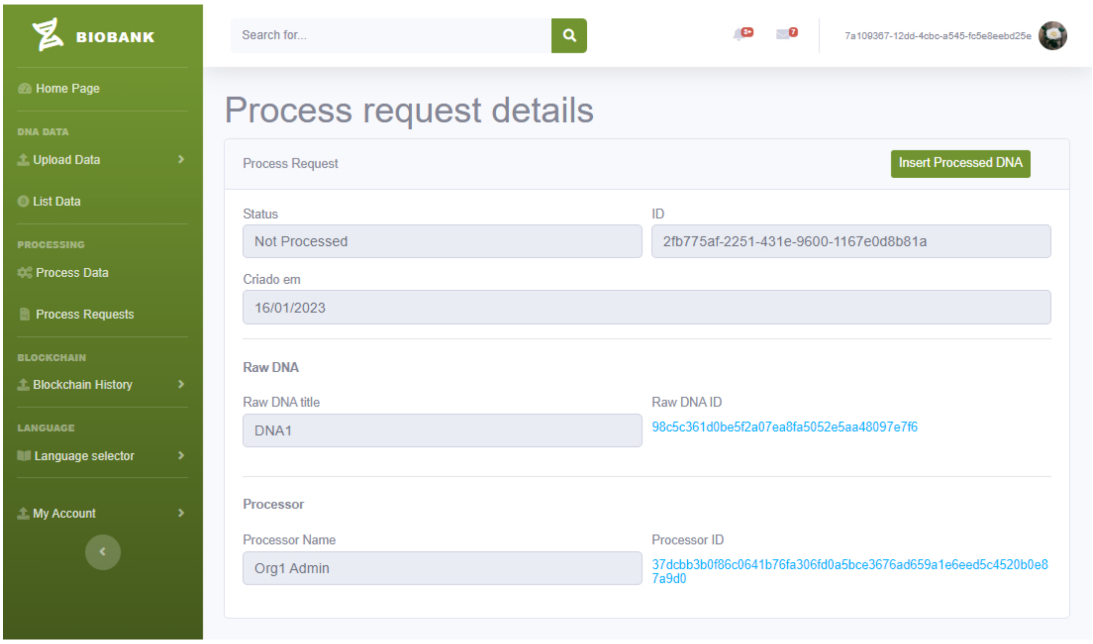
Source: Author

You will use this decryption key to decrypt the DNA data in the Torrente application

A.5 Process DNA

1. Login in biobank-app (possibly using another account)
2. On the home page, click on "process DNA"
3. Choose the DNA data to process. Click on the "process icon"
4. After seeing the DNA data, click on next
5. After seeing the smart contract details (including the payment distribution reserved to Processors), click on "Process"
6. A process request will be created. This is a declaration that you intend to work on this specific raw DNA data (Figure 28)

Figure 28 – Process request details. This process request proves that you intend to work on this data



Process request details

Process Request

Status: Not Processed

ID: 2fb775af-2251-431e-9600-1167e0d8b81a

Criado em: 16/01/2023

Raw DNA

Raw DNA title: DNA1

Raw DNA ID: 98c5c361d0be5f2a07ea8fa5052e5aa48097e7f6

Processor

Processor Name: Org1 Admin

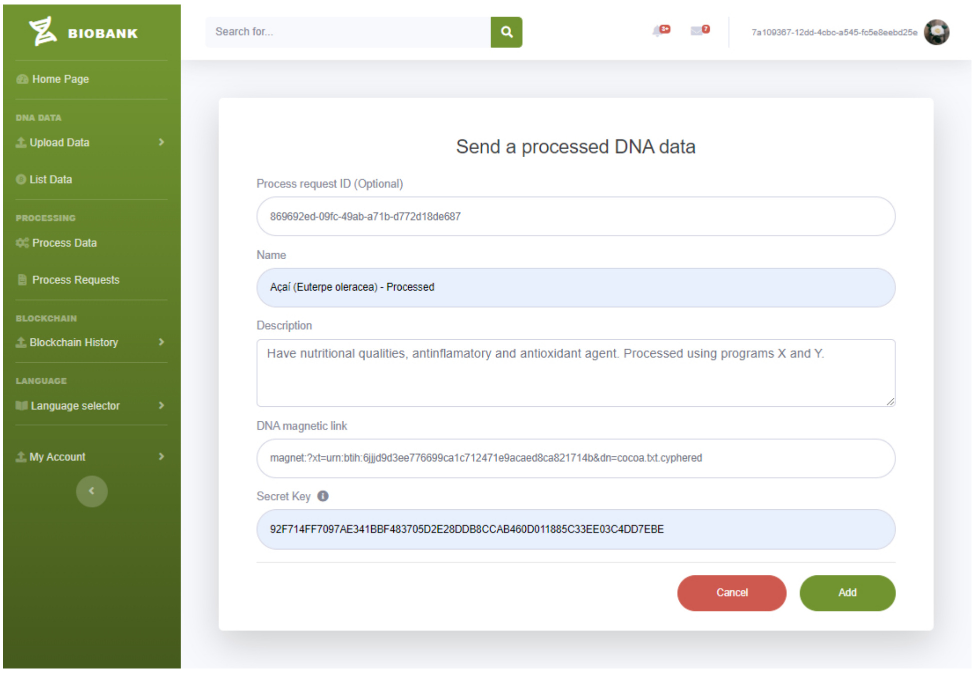
Processor ID: 37dcbb3b0f86c0641b76fa306fd0a5bce3676ad659a1e6eed5c4520b0e87a9d0

Source: Author

DNA data processing is out of the scope of this work. To do this, interested users may employ third-party tools (e.g., nanopore). We will now describe how to insert the processing results

7. On the "processing request details", click on "insert processed DNA".
8. Insert details about the processed data on the form. DNA magnet link and the secret key can be obtained through Torrente (Figure 29)

Figure 29 – Inserting a Process DNA data



Send a processed DNA data

Process request ID (Optional)
869692ed-09fc-49ab-a71b-d772d18de687

Name
Açaí (Euterpe oleracea) - Processed

Description
Have nutritional qualities, antinflamatory and antioxidant agent. Processed using programs X and Y.

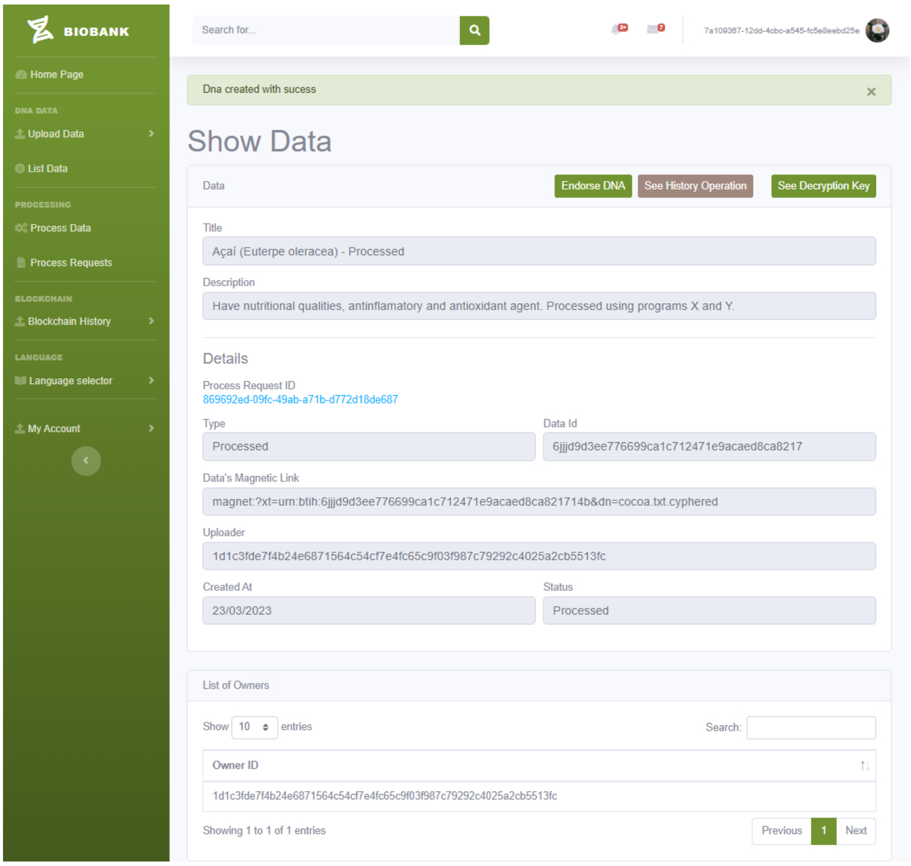
DNA magnetic link
magnet:?xt=urn:btih:6jjjd9d3ee776699ca1c712471e9acaed8ca821714b&dn=cocoa.txt.cyphered

Secret Key
92F714FF7097AE341BBF483705D2E28DDB8CCAB460D011885C33EE03C4DD7EBE

Source: Author

9. Click on "create", to create data on the blockchain (Figure 30)

Figure 30 – Details about an inserted process DNA data.



Dna created with sucess

Show Data

Data

Title
Açaí (Euterpe oleracea) - Processed

Description
Have nutritional qualities, antinflamatory and antioxidant agent. Processed using programs X and Y.

Details

Process Request ID
869692ed-09fc-49ab-a71b-d772d18de687

Type
Processed

Data Id
6jjjd9d3ee776699ca1c712471e9acaed8ca8217

Data's Magnetic Link
magnet:?xt=urn:btih:6jjjd9d3ee776699ca1c712471e9acaed8ca821714b&dn=cocoa.txt.cyphered

Uploader
1d1c3fde7f4b24e6871564c54cf7e4fc65c9f03f987c79292c4025a2cb5513fc

Created At
23/03/2023

Status
Processed

List of Owners

Show 10 entries Search:

Owner ID
1d1c3fde7f4b24e6871564c54cf7e4fc65c9f03f987c79292c4025a2cb5513fc

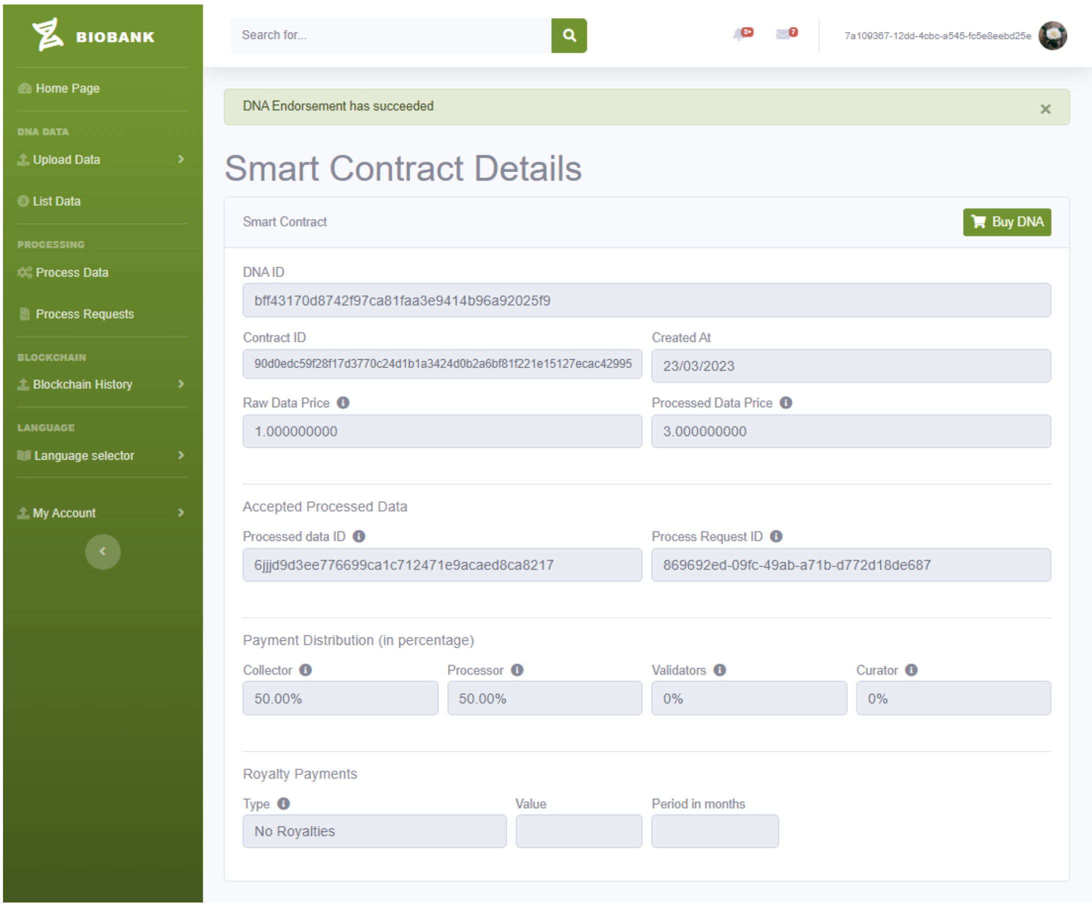
Showing 1 to 1 of 1 entries

Source: Author

10. Click on endorse DNA. This operation submits this "processed data" to be accepted as the "official processed data". As a result, whenever this genetic data is bought, the Processor receives the payment share as described in "payment distribution". If another "processed data" is already accepted as "official", this endorsement will fail.

11. If the endorsement is approved, your data will be included as "official". Note the "accepted processed data" field, which contains the processed data ID and other details.

Figure 31 – Smart contract details, updated with the "accepted processed data"



DNA Endorsement has succeeded

Smart Contract Details

Smart Contract

DNA ID: bff43170d8742f97ca81faa3e9414b96a92025f9

Contract ID: 90d0edc59f28f17d3770c24d1b1a3424d0b2a6bf81f221e15127ecac42995

Created At: 23/03/2023

Raw Data Price: 1.000000000

Processed Data Price: 3.000000000

Accepted Processed Data

Processed data ID: 6jjjd9d3ee776699ca1c712471e9acaed8ca8217

Process Request ID: 869692ed-09fc-49ab-a71b-d772d18de687

Payment Distribution (in percentage)

Collector: 50.00%

Processor: 50.00%

Validators: 0%

Curator: 0%

Royalty Payments

Type: No Royalties

Value:

Period in months:

Source: Author