Functional & Integrative Genomics (2023) 23:101
https://doi.org/10.1007/s10142-023-01015-1

ORIGINAL ARTICLE

Amazon Biobank: a collaborative genetic database for bioeconomy development

Leonardo T. Kimura[1] · Ewerton R. Andrade[1,2] · Ismael Nobre[3] · Carlos A. Nobre[3] · Bruno A. S. de Medeiros[4] · Diego M. Riaño-Pachón[1] · Felipe K. Shiraishi[1] · Tereza C. M. B. Carvalho[1] · Marcos A. Simplicio Jr.[1]

Received: 19 September 2022 / Revised: 12 December 2022 / Accepted: 6 March 2023 / Published online: 25 March 2023

Abstract

Biodiversity 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, it is challenging to adequately compensate non-academic actors such as local populations for their contribution to the innovation process, preventing better bioeconomy development. Although many repositories register genomic data to support biodiversity research, they do not facilitate the fair sharing of economic benefits. In this work, we propose the creation of the Amazon Biobank, a community-based genetic database. We employed blockchain to build a transparent and verifiable log of transactions involving genomic data, and we used smart contracts to implement an internal monetary system for all participants who collect, insert, process, store, and validate genomic data. We also used peer-to-peer solutions to allow users with commodity computers to collaborate with the storage and distribution of DNA files. By combining emerging technologies, Amazon Biobank provides adequate benefit-sharing among all participants that collaborate with data, knowledge, and computational resources. It also provides traceability and auditability, allowing easy association between biotechnological research and DNA data. In addition, the solution is highly scalable and less dependent on the trust deposited in any system player. Therefore, Amazon Biobank can become an important stepping stone to unlock the potential of bioeconomy in rich ecosystems such as the Amazon Rainforest.

Keywords Biodiversity · DNA database · Benefit-sharing · DNA sequencing · Amazon Sustainable Development · Blockchain

Introduction

A biodiversity-based economy, or bioeconomy (Staffas et al. 2013), has been proposed as an alternative for the sustainable economic development of regions with high biological diversity, such as Amazonia (Nobre et al. 2016; Nobre and Nobre 2019). Amazonian ecosystems are home to 25% of the world's tree species (Beech et al. 2017; ter Steege et al. 2013) and a comparable share in organisms such as microbes, fungi, and animals. This incredible biodiversity

✉ Marcos A. Simplicio Jr.
mjunior@larc.usp.br

Leonardo T. Kimura
lkimura@larc.usp.br

Ewerton R. Andrade
ewerton.andrade@unir.br

Ismael Nobre
nobreismael@gmail.com

Carlos A. Nobre
cnobre.res@gmail.com

Bruno A. S. de Medeiros
bdemedeiros@fieldmuseum.org

Diego M. Riaño-Pachón
diego.riano@cena.usp.br

Felipe K. Shiraishi
fkspoli@usp.br

Tereza C. M. B. Carvalho
terezacarvalho@usp.br

[1] University of Sao Paulo, Sao Paulo, SP, Brazil

[2] Federal University of Rondonia, Porto Velho, RO, Brazil

[3] Amazon 4.0 Institute, Manaus, AM, Brazil

[4] Field Museum, Chicago, IL, USA

evolved throughout more than 50 million years (Hoorn et al. 2010), providing critical ecological services whose annual value can reach trillions of dollars (Strand et al. 2018). Moreover, human populations have lived in the Amazon region for thousands of years, having a historical impact on its landscape, domesticating species, and holding a large body of knowledge on the management and usage of its biodiversity (Levis et al., 2018). Both the outstanding diversity of the region and the traditional knowledge about it provide a huge potential to foster local development by innovation via biomimetic engineering, synthetic biology, and the development of new materials, chemical compounds, and biofuels (Nobre and Nobre 2019; Rech 2011).

Albeit promising, major challenges still exist to unlock the potential of bioeconomy in the Amazon region. One of them is the scale of research effort necessary to survey millions of species across more than 5 million square kilometers of forest. Another challenge resides in enabling broad access to data and products derived from Amazonian organisms while ensuring regulatory compliance and fair compensation of local populations for their traditional knowledge, conservation efforts, and participation in innovation process. These challenges and opportunities are particularly salient in the field of biodiversity genomics. Together with other sources of data, genomics can be an invaluable tool in conservation (Watsa et al., 2020), biodiversity monitoring (Hobern and Hebert 2019), tracing of natural products (Ng et al. 2020), discovery of functional molecules (Toyama et al. 2018), crop management (Boykin et al. 2018), and monitoring of human and wildlife pathogens (Deng et al. 2020). Recent developments in low-cost, portable DNA sequencing hold promise for a larger user base of genomic methods, including local populations of high-biodiversity regions (Watsa et al. 2020). However, the field has been historically operated under a colonialist model (Li 2021). Beyond sequencing technology, new practices and methods of data sharing are key to turn the potential of biodiversity genomics into local development. One step is the construction of highly scalable genomic databases that can be populated by residents of areas of interest, who would then preserve ownership of the collected data and be compensated accordingly.

Today, many repositories register genomic data for supporting biotechnological research, both for academic and industrial purposes. Because sequencing comparison is the basis of most of our understanding of genomics, these repositories are fundamental tools. The most important general-purpose repositories include those supported by the US National Center for Biotechnology Information (NCBI) (Sayers et al. 2020), by the European Bioinformatics Institute (EBI) in Europe (Harrison et al. 2021), and by the DNA Databank of Japan (DDBJ) (Fukuda et al. 2021). They are all situated in developed countries and make data publicly available, without usage tracing. Such a model may make data re-usage easier, but one downside is that it also discourages data sharing among producers, at least until the data is already published. This can have harmful consequences when rapid data sharing is required, such as in tracking epidemics, which is leading to the development of new models. For example, GISAID (Elbe and Merrett 2017) is a specialized database for genomic data on respiratory viruses that requires registration and adherence to a code of conduct for data access, fostering academic collaboration.

Despite the relevance of such repositories as clusters of genomic data, one challenge that remains is that they do not facilitate adequate sharing of the economic benefits resulting from their services. Moreover, they do not provide a platform for collaborative data processing and sharing of results. With portable low-cost DNA sequencing, there is an opportunity for new actors beyond academia to engage in genomic sequencing. However, data sharing among this extended user base still lacks incentives.

Aiming to tackle this issue, this work describes the 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, 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 and Nobre 2019). Managed by the Amazon 4.0 Institute (Amazonia 4.0 Institute 2022), these laboratories seek to prepare and train local communities into exploring high-value bioeconomy opportunities. One of those opportunities consists in acting as data collectors 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.

Materials and methods

The main goal of the Biobank is to promote the collaborative development of biotechnology research in regions with rich ecosystems. The system is designed to handle data from environmental and biodiversity origin, excluding human DNA. For this purpose, some functionalities are considered essential requirements:

- 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., through 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 be 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. In particular, 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 results in better 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 data in the Biobank when handling disputes involving data misuse. Also, regarding 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.

Blockchain

A blockchain is essentially a structure for storing data in blocks, forming a chronological chain where each block is cryptographically linked to the previous one via hash functions (MIT, 2018). As a result, knowing one block is enough to check whether a previous block is also part of the chain or if it has been tampered with. This leads to a transparent and reliable data log. Commonly, blockchains are implemented in a distributed manner, with many entities keeping identical copies of its contents and using a consensus protocol to ensure consistency (Swan 2015).

One typical usage of blockchains refers to the exchange of digital assets among participants that do not trust each other. This is the case, for example, of Bitcoin (Nakamoto 2008), a digital currency used worldwide without any central authority for managing its operation. Bitcoin uses blockchain as a public ledger containing all transactions validated by its peers, also called miners, after verifying that those transactions are digitally signed by the sender and have enough funds to be executed. For this validation process and for participating in the underlying consensus protocol that orders those transactions, miners are rewarded with bitcoins.

While Bitcoin adopts a highly open and decentralized model for its operation, many alternative blockchain-based systems impose some restrictions on who can read from the blockchain, write into it, or have permission to participate in its consensus protocol (Shrivas 2019; Zutshi et al, 2021). Usually, restrictions on reading capabilities are employed to facilitate privacy protection among blockchain participants, even if the blockchain stores sensitive data in plaintext; however, this approach does not provide privacy guarantees toward the entities that control the reading permissions. Restrictions on writing capabilities, in turn, are usually coupled with an identification management system aiming to dissuade misbehavior, since peers engaged in malicious activities can be suspended or permanently evicted from the system. Finally, specifying that only a small number of well-defined entities can participate in the consensus protocol, creating a permissioned blockchain, usually allows for more lightweight consensus mechanisms, like Raft (Ongaro and Ousterhout 2014) or Consensus (Chase and MacBrough 2018).

In all cases, independent auditors who monitor blockchain updates can ensure the transparency and reliability of the system entries: given the cryptographic link between any block and its predecessors, this is enough to detect attempts to modify data already registered in the blockchain, if such attempts ever occur (Laurie 2014).

In the context of the Amazon Biobank, the choice of a blockchain as part of the architecture 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 "Register DNA sequences", this ensures, for example, that a user who uploads some DNA data into the Biobank cannot have this data associated with another user, even if the latter colludes with malicious system entities. Analogously, when the platform is paid for some service, users can verify that the total amount of biocoins involved in the transaction is shared among the appropriate players — or, at least, that attempts to do otherwise can be easily detected, 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, it is quite naturally obtained with the help of 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, it is somewhat natural to use a permissioned blockchain with a lightweight consensus mechanism. Our prototype adopts a Raft-based mechanism, which can add a new block in less than 1 s and supports at least two hundred transactions per second (Wang and Chu 2020).

Furthermore, since the database does not store any sensitive data (only magnet links for encrypted data, as further discussed in "BitTorrent"), 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).

BitTorrent

Another important component of the Biobank is its underlying data distribution technology: BitTorrent (Cohen 2003). In a nutshell, BitTorrent is a highly scalable P2P protocol for data sharing, by means of which users can simultaneously download and upload pieces of a file. This approach allows servers to partially shift the burden of file hosting toward users who keep downloaded files in their own local storage, becoming content "seeders". At the same time, BitTorrent associates each data piece to its hash value in a torrent file that unequivocally represents the complete data content, thus providing strong data integrity guarantees. It also supports distributed discovery and download of torrent files, using a distributed hash table (DHT) where the search keys, essentially 20 bytes hashes, are named "magnet links" (Wolchok & Halderman 2010). In our solution, the blockchain only stores such magnet links, thus ensuring a small block size despite the size of the corresponding DNA data (Fig. 1).

The adoption of BitTorrent for data sharing in the Amazon Biobank is important for creating a more scalable and less centralized solution, when even users with commodity computers can collaborate with the storage and distribution of DNA files. The need for such characteristics in the target scenario becomes evident when we consider the frequently large sizes of DNA files. Hence, although the system managers may decide to keep a copy of all DNA files for ensuring their long-term availability, the resulting storage service can be made more affordable whilst still highly available and redundant with the support from the system users. To encourage users' collaboration as seeders, we combine BitTorrent with the Torrente micro-payment mechanism (Shiraishi et al. 2021), similarly to decentralized storage solutions like Filecoin (Benet et al. 2017) and BitTorrent Speed (BitTorrent Foundation 2019). Meanwhile, to avoid unauthorized distribution of DNA data by peers, all content is stored in encrypted form. Consequently, although seeders collaborate by storing and distributing data, only users who obtain the decryption key (e.g., by purchasing it) can access the corresponding plaintext.

Table 1 Blockchain permissions to perform different operations in Amazon Biobank

Operation	Allowed users
Write	Identified users
Read/audit	Any users
Consensus participation	Only system managers

Related works

Several studies discuss the benefits of blockchain for registering and distributing DNA data. For instance, Ozercan et al. (2018) and Thiebes et al. (2020) discuss the integration of blockchain and distributed sharing techniques to create a scalable and privacy-preserving genomic data distribution architecture. Likewise, Alghazwi et al. (2021) discuss that successful solutions must provide (1) data redundancy, by identifying and eliminating duplicated data and (2) data verifiability, meaning that the accuracy of distributed processed tasks done by untrusted nodes should be somehow validated. To address these challenges, Amazon Biobank employs incentive mechanisms to motivate the participation of data validators, besides employing a combination of curatorship, data processing validation, and a reputation system.

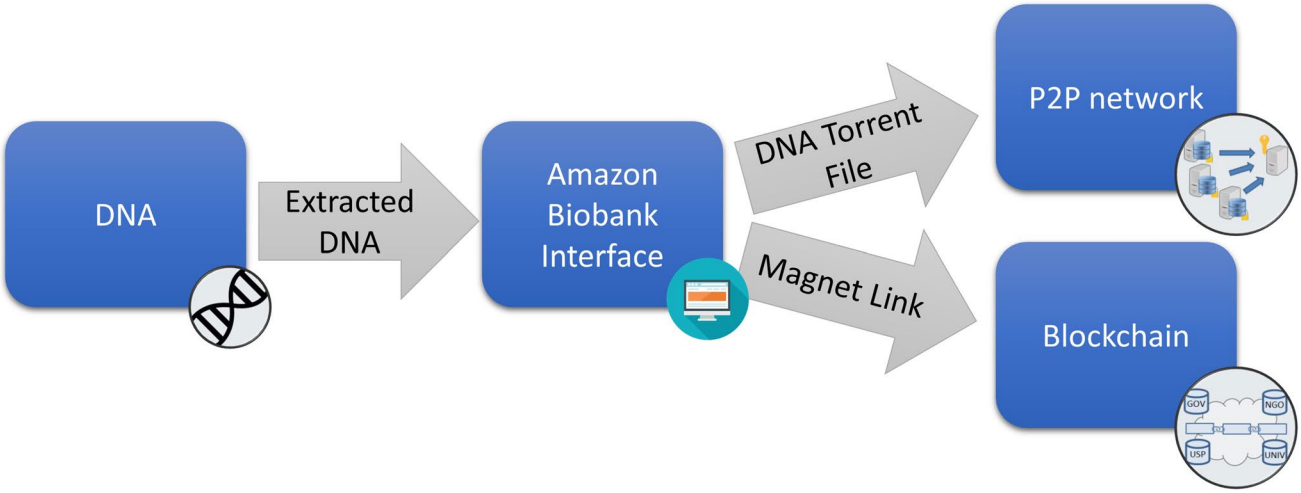

Fig. 1 High-level steps to upload DNA data in Amazon Biobank. Firstly, the extracted DNA is inserted in the Amazon Biobank Interface. Then, the DNA torrent file is stored in a P2P network, and only a lightweight magnetic link is registered on the blockchain

Besides such academic studies, there are biotechnology companies like Nebula Genomics (Grishin et al. 2018) and Genesy (Carlini et al. 2019) that also employ blockchain in their business models. Specifically, they encourage people to share their genomic data by (1) providing easy access to sequencing services and (2) rewarding collaborators with an internal cryptocurrency. However, those solutions prioritize human genetics. Hence, they give little emphasis on biodiversity as an asset, or on biotechnology as a service to be built on top of their platform. They also do not implement intellectual protection or royalty's payments; features considered an important requirement for promoting collaboration.

Among initiatives focused on biodiversity and on implementing the Nagoya Protocol (Buck and Hamilton 2011), there are calls for multilateral agreements for benefit-sharing that use the current data infrastructure (Scholz et al. 2022). The relevant actors, in this case, are countries instead of individuals or institutions generating and processing data, being complementary to our approach. Additionally, the United Nations Development Programme (UNDP) is conducting a blockchain-based project to improve genetic resource traceability and benefit-sharing (UNDP 2021). However, the system focuses on natural products, like plants or natural substances, not on genetic data itself. Unlike our work, none of them handles collaborative storage and distribution of genomic data. Finally, both proposals require global coordination between countries, which is still a work in progress

System architecture

In this section, we discuss the design decisions taken for fulfilling the requirements described in "Materials and methods". We start by describing the main roles played by the entities that compose the system (see Fig. 2), and then describe in detail how they interact to enable the system's operation.

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.

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.

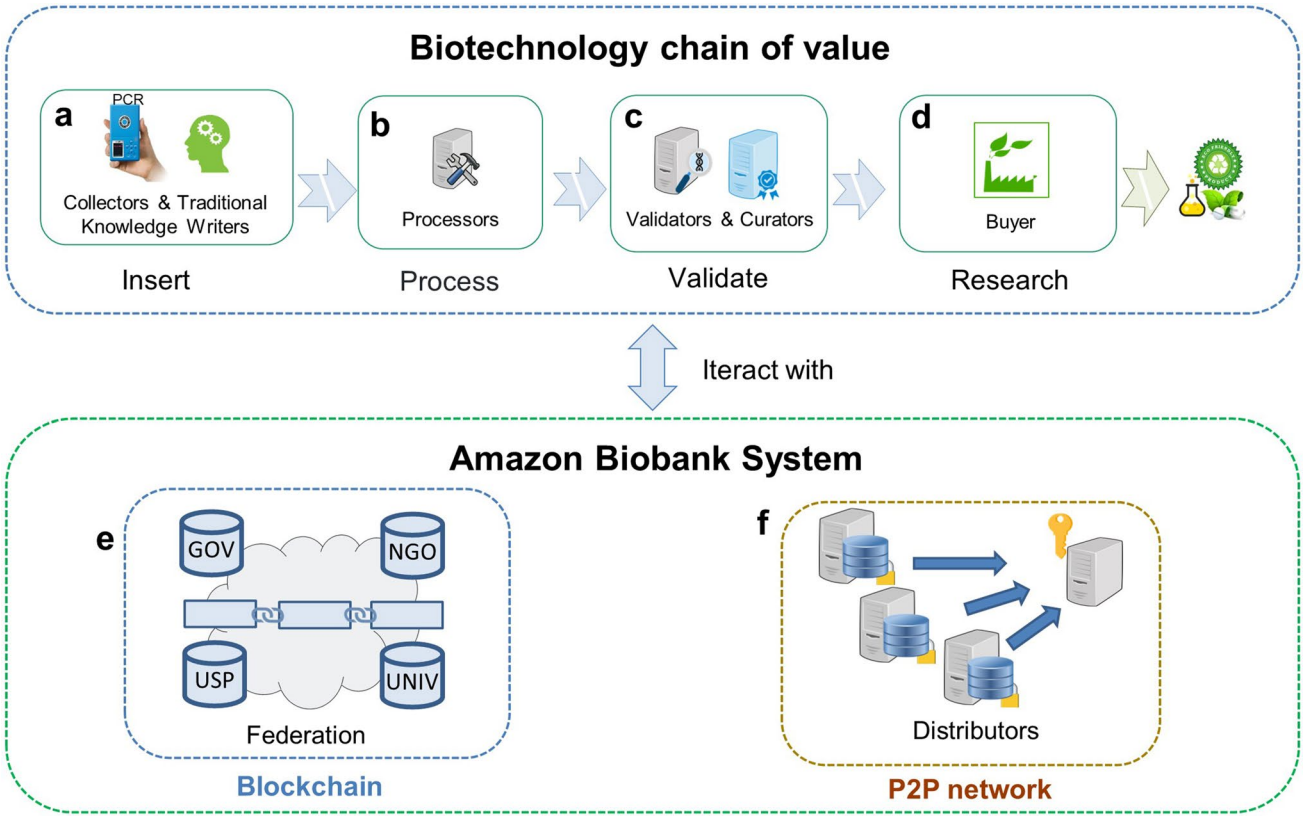

Fig. 2 Amazon Biobank: overview and main operations. (**a**) Collectors provide raw DNA data, possibly complemented with traditional knowledge files. (**b**) Processors sequence and assemble raw DNA. (**c**) Validators and curators assess data correctness. (**d**) Buyers purchase access to DNA data, using it for their research. (**e**) The Federation comprises universities and (non-)governmental organizations who manage the system. (**f**) Distributors store and share data over a BitTorrent-based P2P network

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.

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.

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 in 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.

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.

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 act when a given data entry receives multiple positive and negative votes by validators, or when someone claims that a raw DNA entry is bogus.

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 et al, 2018).

Register DNA sequences

One of the Biobank system's primary operations is the registration of raw DNA data (see Fig. 3). 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

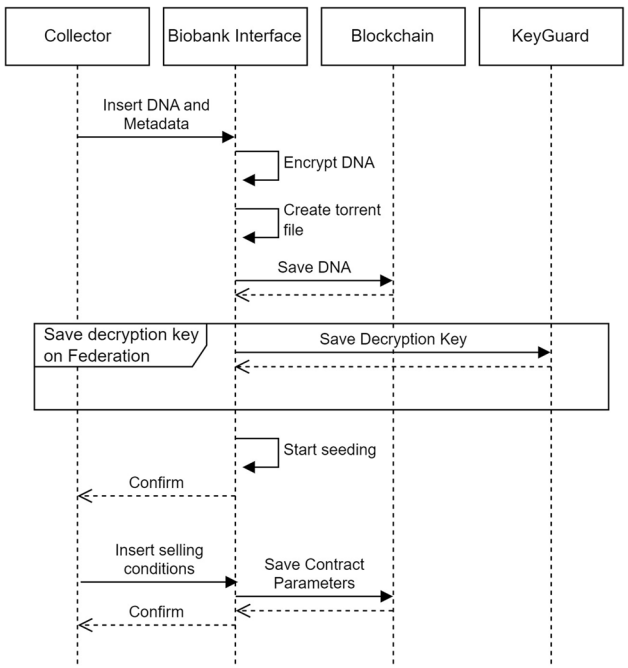

Fig. 3 Sequence diagram — upload DNA data

Inserting genetic data

A collector (e.g., a resident of the Amazon region) collects some DNA sequences using a portable sequencing device. The result may be either a raw instrument signal or a DNA sequence read, depending on the type of equipment used.

The collector then encrypts the collected data, associates it with any relevant metadata, and creates the corresponding torrent file. The latter is then uploaded to the system to be associated with the collector, assuming the magnet link for the same torrent file does not already exist. Once this association is registered in the blockchain, the collector can be confident about its ownership, i.e., that the collected data cannot be "hijacked" by some other user in collusion with system administrators. Thereafter, the collector can share the encrypted data with distributors, who can help with the storage and delivery of data pieces to interested parties. By default, the collector is also expected to share the decryption key with the federation, facilitating the implementation of services like sequence search, as well as data availability control. Nevertheless, collectors having enough computational capabilities may prefer to keep the decryption key to themselves, and then handle search and/or backup services on their own.

As aforementioned, when creating the torrent file, the collector is expected to associate any useful information with 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, 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), among others. 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.

Verifying data correctness

The correctness of the inserted metadata is important to avoid situations in which buyers end up purchasing access rights to incorrect data (e.g., randomly generated data or data that mismatches its description). Even though mistakes are not completely avoidable, the Biobank promotes confidence in the registered data by combining two mechanisms. The first is curatorship: players registered as curators (typically biologists) can verify the correctness of the registered metadata, besides inserting their own. For that purpose, curators must follow verification guidelines published and periodically reviewed by the federation. When one or (preferably) more curators provide their stamp of approval upon an entry and crosscheck their contributions, the interest by potential buyers on the corresponding file is prone to increase. The second is a reputation system, according to which players who provide correct metadata are rewarded with a higher reputation and are penalized otherwise. Users with low reputations may have a part of their funds withheld as a stake, enabling the system to compensate peers affected by their misbehavior, besides rewarding those responsible for detecting the misdeed. In the worst-case scenario, misbehaving users may be evicted from the system, losing access to their funds. As a result, besides curatorship, the quality of the registered information may also be assessed indirectly, via the reputation of the corresponding collector. This combination of verifications, together with some automatic validation mechanisms for assessing genome assembly and annotation completeness (Seppey, Manni, & Zdobnov, 2019), should dissuade malicious activities such as the generation and upload of false DNA sequences resembling real data.

In addition, to encourage diversity of data inserted in the system, the Biobank itself can mint biocoins and reward them to collectors for DNA samples evaluated as unique. 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). On the other hand, this rewarding mechanism favors collectors who provide more value-adding information. At the same time, it limits the incentives for users who decide to simply provide readings for the same specimen repeatedly, polluting the system with redundant information. The responsibility for evaluating uniqueness may involve both: (1) automated mechanisms enabling similarity comparisons with entries in the Biobank and public databases (Goldstein & DeSalle 2011) and (2) manual intervention by curators. This task may also identify attempts to insert duplicated data in the Biobank, triggering adequate actions (e.g., decreasing the reputation of users who engaged in pollution attacks).

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. Alternatively, and whenever desired, collectors may decide to explore the data themselves, restricting its access to a limited number of players. In this case, the proposed platform would still be useful as a Timestamp Server, enabling data owners to register their intellectual property and to exchange knowledge. 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 limits. For example, 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, which is among the reasons why the system entities are also rewarded when buyers purchase access rights to data, and collectors in particular receive freshly minted coins for unique data.

Associated 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" (IGC 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 "Register DNA sequences" also apply to TK, including the registration of smart contracts that define pricing and royalties' rules. However, very specific regulations may apply in this case. 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. The same law also covers the need

for additional contracts involving free, prior, and informed consent (FPIC). Whenever technically possible, the system can enforce such regulations using suitable smart contracts. For example, in the above example, biocoins received for the rights to access the corresponding TK could be sent to an account registered for the entire community or government. In the case of FPIC requirements, the Amazon Biobank can transparently store required authorizations. Other rules can be flexibly encoded, and purchasing this registry implies that the buyer has accepted those rules. Therefore, the system 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.

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 "BitTorrent", the 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, the motivation for doing so is that distributors can be paid by the corresponding downloaders. For example, raw DNA readings can be downloaded by users playing the role of processors, whereas traditional knowledge and processed DNA data would be of interest to buyers. For each of the (possibly many) data pieces sent to the downloader, the distributor is expected to receive some fixed amount of biocoins as a reward. 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), which can be seen as an optimized version of the protocol employed by BitTorrent Speed (BitTorrent Foundation 2019). Essentially, this means that multiple small payments can be made to the distributor by the downloader, and then be registered in batches in the blockchain, using a single transaction for the total amount due. Note that downloaders can themselves decide to become distributors after they receive data pieces, and then be similarly rewarded for their contribution to the system.

This collaborative, BitTorrent-based distribution approach creates an infrastructure with high scalability and availability. Even though this reduces the burden on the nodes that form the federation, they should still participate in some procedures related to the storage and distribution of DNA data. In particular, the federation is expected to facilitate the task of finding BitTorrent nodes and contents, besides monitoring the network's health. For this purpose, the federation should maintain an active BitTorrent tracker and also take part in BitTorrent's DHT. In addition, federation nodes should act as seeders for newly registered content, possibly together with the collector who uploaded it. This procedure facilitates the initial data distribution and may involve reduced (or even no) fees for interested downloaders. Similarly, whenever the number of distributors available for some content is too small, federation nodes can once again assume the role of seeders, ensuring long-term availability of the Biobank contents. Since popular contents are unlikely to require constant seeding by the federation, though, keeping them in a cold storage facility may be a suitable and cost-effective approach.

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, as well as any relevant metadata (e.g., the scripts and programs employed), are also turned into a BitTorrent file and added to the system in the form of processed DNA. To obtain the corresponding raw data, though, processors (like any other entity) in principle need to spend some biocoins to pay distributors for the downloaded data pieces and collectors for the corresponding decryption key. In addition, 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. In practice, the main reward is the fact that processors participate in the distribution of royalties for the DNA data they register in the system. Nevertheless, since predicting the actual interest of a given entry in the Biobank is hard, 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 resulting processed data is registered in the blockchain.

Due to the distributed and collaborative nature of the Biobank, more than one processor might work on the same DNA sequence simultaneously, using the same settings. If not correctly handled, this can lead to very similar assembled or annotated DNA sequences being registered to different processors. To avoid unnecessary 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). The computational efforts of processors whose results are obtained at a later time are not wasted, though. 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 et al. 2019), this approach promotes better data quality and trustworthiness. For example, entries approved by several 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.

Search

The Biobank can support a few types of search procedures, based on (1) plaintext keywords, (2) encrypted keywords, and (3) DNA sequences.

Keyword searches are performed on the metadata registered in the system. When the corresponding metadata is stored as plaintext, this procedure can be as simple as a regular keyword search in the database maintained by the federation; to facilitate distribution, it may also involve searches in a DHT where such keywords are stored. Handling distributed searches over encrypted data is also possible if the data owners are willing to deploy searchable-encryption mechanisms; it should be noted, though, that doing so usually leads to a probabilistic amount of information being leaked with each search (Agarwal and Kamara 2019).

Finally, it is possible to support sequence-based searches using BLAST (Altschul et al. 1990) or similar methods. For that purpose, the nodes of the Federation may coordinate the search on their locally stored DNA data, assuming that those nodes have access to the corresponding decryption keys (which is the default case). Also, given the higher computational costs of such search operations, the fees imposed on buyers in this case may be steeper than those involved in simple keyword searches.

Once again, other nodes from the network can collaborate with this task, as long as they themselves have access to non-encrypted DNA data. For example, buyers who have previously acquired access to the DNA data may want to recover part of their investment, by sharing part of the corresponding search fees. Collectors, processors, and validators, on the other hand, may engage in such activities not only to receive those fees, but also because it is in their best interest to increase the visibility and selling potential of their own data. In such cases, the collaborating nodes looking for correspondences on local data return the results to the federation, which validates the responses before conveying them to the requester. Invalid results, like reporting a non-existing correspondence, lead to penalties to the misbehaving entities.

Finally, it is worth noting that requesters may prefer to disallow collaborative searches, and instead keep the queries inside the federation. After all, the information about what is being searched may be considered strategic or sensitive, in special when the search involves DNA sequences. Therefore, even though collaborative search procedures may reduce the processing costs and response latency for each query, requesters should be allowed to disable this feature.

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 keys 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. That purchase is registered in the blockchain and processed via smart contracts that define the distribution of the biocoins, as discussed in "Register DNA sequences."

The same smart contract also defines how royalties for products generated from the purchased data should be paid and distributed. The blockchain can then be used as a transparent log to trace the origin of such biotechnology products to the Amazon Biobank. Obviously, though, the existence of such a contract does not prevent dishonest buyers from using the acquired data without giving credit to Amazon Biobank or honoring the corresponding payments. Nevertheless, this 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 in the 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.

As another possibility, buyers can also request data access for a specific organism, and then remunerate whoever provides the DNA with that specification. 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.

Discussion

The 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 all requirements described in "Materials and methods":

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 the confidence on 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 nonfunctional 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. Also, the blockchain stores only a small reference to DNA data, employing a permissioned consensus that allows high-throughput and low-latency transactions. The resulting collaborative architecture is, thus, expected to be capable of handling many users and data.

Remaining challenges

The proposed architecture for the Amazon Biobank addresses many of the challenges faced by (collaborative) genetic repositories. However, some of the issues discussed in Ito & O'Dair (2019) remain as open questions, in special: how to ensure the correctness of inserted data and the provenance problem, i.e., how to avoid transfer of ownership outside the network. In both cases, what the Amazon

Biobank provides are mechanisms to attenuate, rather than solve, these problems.

First, if some raw DNA reading or sequenced DNA is crafted with the specific purpose of resembling real data, it would be hard to use automated tools to detect the deceit. Actually, data readings from the same sample are likely to be different from each other, making it easy for dishonest collectors to fill the system with redundant data. To dissuade such misbehavior, the system does not automatically remunerate collectors for just any piece of data. Instead, rewards are provided only after the raw DNA entry is processed by independent processors, who validate each other's results, and (possibly manually) assessed by curators. On the other hand, when misbehavior is detected, users (1) lose reputation, which is prone to reduce the interest on the entries owned by them and (2) may be suspended or evicted from the system, losing access to existing funds. Therefore, this approach is expected to create a scenario where users have more to lose than gain from misbehaving.

Second, it is evident that the traceability provided by blockchains is not ubiquitous. This means, for example, that entities in possession of plaintext data obtained from the Biobank may decide to distribute this data freely (e.g., over a regular BitTorrent network). They may also sell it directly to interested buyers, keeping the monetary gains for themselves instead of sharing profits with other system players. Likewise, after obtaining the desired data (either from the Biobank or via an illegal channel), users may decide not to register the resulting products/services as "Biobank-based," aiming to avoid the associated fees or royalties. Arguably, though, the benefits provided by using and referencing the Amazon Biobank as a source of data should at least dissuade such malicious actions. After all, as discussed in "Purchasing access rights to data", data traceability is usually an important requirement in scenarios where the Biobank products will be used (e.g., biotechnology research).

Therefore, obtaining the data via dubious channels is not in the best interest of potential buyers. Then, after the data is acquired, exposing it to third parties would facilitate competition, especially in the case of industry players. Those same players are also likely to be interested in referencing the Amazon Biobank as a source of new products/services as part of ESG-oriented marketing campaigns. Besides, whenever the undue usage of Biobank data is observed in the wild, the Biobank's traceability would still be useful for resolving disputes: the system's underlying blockchain can give an approximate time frame of when the DNA sequence appeared in the Biobank. By comparing this date with the release of products/services that are suspected to be derived from such data, and analyzing the data itself, it would be easier to produce evidence to support violation claims.

Once the Amazon Biobank reaches production phase, its biocoins may be subject to financial regulations from different countries. Therefore, some adjustments may be necessary to enforce compliance. For example, it may be necessary to restrict biocoins trading only through certified and regulated cryptocurrency exchanges (e.g., Binance, Coinbase, or Kraken).

Finally, the proposed system is designed to simply handle biodiversity data, making no claims over copyright or restricting the usage of that data. This is similar to practices adopted by existing systems such as the NCBI (Sayers et al. 2020). Moreover, and as aforementioned, 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. Analogous challenges may also arise from handling traditional knowledge associated with genomic data from biodiversity. Even though such data it is not strictly necessary for the functioning of the Biobank, we envision that registering TK in an encrypted form may be interesting for traditional knowledge holders. At least, this 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 accesses to this data, if allowed, can then be mediated using smart contracts, and generate verifiable records in the blockchain. 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.

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 system player.

We currently have a working prototype of the proposed system (Kimura, Andrade, Carvalho, & Junior, 2021), with which it is possible to demonstrate the entire life cycle of genetic data (collection, registration, processing, and purchase). The next steps planned include deploying this prototype in the Amazon region, still as a small-scale demonstration, in collaboration with the Amazon 4.0 initiative. After those initial tests, the next steps include a larger-scale assessment and 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 2017). If successful, we believe that the Amazon Biobank can be an important stepping-stone to unlocking the huge potential of bioeconomy in rich ecosystems such as 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.

Abbreviations ESG: Environmental, social, and governance; DHT: Distributed hash table; P2P: Peer-to-peer; TK: Traditional knowledge

Author contribution Ismael Nobre and Carlos Nobre conceived the idea. Marcos Simplicio, Tereza Carvalho, Bruno Medeiros, Diego Riaño-Pachón, Ewerton R. Andrade, and Leonardo Kimura designed the architecture. Leonardo Kimura and Felipe Shiraishi implemented the prototype. Leonardo Kimura, Marcos Simplicio, Bruno Medeiros, and Diego Riaño-Pachón wrote the main manuscript text and prepared all figures and tables. All the authors revised the manuscript. The author(s) read and approved the final manuscript.

Funding This work was supported by Ripple's University Blockchain Research Initiative, and in part by the Brazilian CNPq (grants 304643/20203 and 310080/2018–5).

Data availability The Amazon Biobank prototype is available in https://github.com/amazon-biobank/biobank-FIG, including the documentation about the installation, basic requirements, and some performance numbers.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Agarwal A, Kamara S (2019) Encrypted distributed hash tables. Cryptology ePrint Archive, Report 2019/1126

Alghazwi M, Turkmen F, van der Velde J, Karastoyanova D (2021) Blockchain for genomics: a systematic literature review. Retrieved from https://doi.org/10.1145/3563044

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment Search tool. J Mol Biol 215(3):403–410

Amazonia 4.0 Institute (2022) Homepage. https://amazonia4.org/. (Accessed in 18–08–2022)

Bach LM, Mihaljevic B, Zagar M (2018) Comparative analysis of blockchain consensus algorithms. *41st* Int. ConventiInfo Commun Technol, Electron Microelectron, 1545– 1550

Beech E, Rivers M, Oldfield S, Smith PP (2017) GlobalTreeSearch: The first complete global database of tree species and country distributions. J Sustain for 36(5):454–489

Benet J, Dalrymple D, Greco N (2017) Proof of replication. Protocol Labs, July, 27, 20

Boykin L, Ghalab A, De Marchi BR, Savill A, Wainaina JM, Kinene T, others (2018) Real time portable genome sequencing for global food security. F1000Research, 7(1101), 1101

Buck M, Hamilton C (2011) 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 20(1):47–61

Carlini R, Carlini F, Dalla Palma S, Pareschi R (2019) Genesy: a blockchain-based platform for DNA sequencing. DLT@ ITASEC, 2019, 68–72

Chase B, MacBrough E (2018) Analysis of the XRP ledger consensus protocol. arXiv preprint

Cohen B (2003) Incentives build robustness in BitTorrent. Workshop on economics of peerto-peer systems (Vol. 6, pp. 68–72)

Deng X, Achari A, Federman S, Yu G, Somasekar S, B´artolo I, ..., Chiu CY (2020) Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. Nat Microbiol 5(3):443–454. Retrieved from https://doi.org/10.1038/s41564-019-0637-9

Elbe S, Merrett GB (2017) Data, disease and diplomacy: Gisaid's innovative contribution to global health. Global Challenges 1(1):33–46. Retrieved from https://doi.org/10.1002/gch2.1018

BitTorrent Foundation (2019) BitTorrent speed. https://www.bittorrent.com/token/bittorrent-speed/

Fukuda A, Kodama Y, Mashima J, Fujisawa T, Ogasawara O (2021) DDBJ update: streamlining submission and access of human data. Nucleic Acids Res 49(D1):D71–D75. Retrieved from https://doi.org/10.1093/nar/gkaa982

Goldstein PZ, DeSalle R (2011) Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. BioEssays 33(2):135–147

Grishin D, Obbad K, Estep P, Cifric M, Zhao Y, Church G (2018) Blockchain-enabled genomic data sharing and analysis platform (Tech. Rep.). Miami, Florida: Nebula Genomics

Harrison PW, Ahamed A, Aslam R, Alako BTF, Burgin J, Buso N, ..., Cochrane, G (2021) The European nucleotide archive in 2020. Nucleic Acids Res 49(D1):D82–D85. Retrieved from https://doi.org/10.1093/nar/gkaa1028

Hobern D, Hebert PD (2019) Bioscan: revealing eukaryote diversity, dynamics, and interactions. Biodivers Inf Sci 3 37333. Retrieved from https://doi.org/10.3897/biss.3.37333

Hoorn C, Wesselingh FP, ter Steege H, Bermudez MA, Mora A, Sevink J, ..., Antonelli A (2010) Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. Science 330(6006), 927–931. Retrieved from 110.1126/science.1194585

IGC (2022) Intergovernmental committee on intellectual property and genetic resources, traditional knowledge and folklore. World Intellectual Property Organization - https://www.wipo.int/tk/en/igc/.

Ito K, O'Dair M (2019) A critical examination of the application of blockchain technology to intellectual property management. Business transformation through blockchain (pp. 317–335). Springer

Kimura L, Andrade E, Carvalho T, Junior MS (2021) Amazon Biobank - a community-based genetic database. Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg) (pp. 74–81). Porto Alegre, RS

Laurie B (2014) Certificate transparency. Commun ACM 57(10):40–46

Levis C, Flores BM, Moreira,PA, Luize BG, Alves RP, Franco-Moraes J, ..., Clement CR (2018) How people domesticated amazonian forests. Front Ecol Evol 5:171. Retrieved from https://doi.org/10.3389/fevo.2017.00171

Li F-W (2021) Decolonizing botanical genomics. Nat Plants 7(12):1542–1543. Retrieved from10.1038/s41477–021–01041–6

MIT (2018) A glossary of blockchain jargon. MIT technology review. Retrieved from https://www.technologyreview.com/2018/04/23/143486/a-glossary-of-blockchainjargon/

Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. Decentralized Bus Rev 21260

Ng CH, Ng KKS, Lee SL, Tnah LH, Lee CT, Zakaria NF (2020) A geographical traceability system for Merbau (Intsia palembanica Miq.), an important timber species from peninsular Malaysia. Forensic Sci Int: Genetics 44:102188. Retrieved from https://doi.org/10.1016/j.fsigen.2019.102188

Nobre CA, Sampaio G, Borma LS, CastillaRubio JC, Silva JS, Cardoso M (2016) Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. Proc Natl Acad Sci 113(39):10759–10768. Retrieved from https://doi.org/10.1073/pnas.1605516113

Nobre I, Nobre CA (2019) 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.

Ongaro D, Ousterhout J. (2014). In search of an understandable consensus algorithm. *Usenix annual tech. conf.* (pp. 305–319).

Ozercan HI, Ileri AM, Ayday E, Alkan C (2018) Realizing the potential of blockchain technologies in genomics. Genome Res 28(9):1255–1263

Rech E (2011) Genomics and synthetic biology as a viable option to intensify sustainable use of biodiversity. Nat Preced Retrieved from https://doi.org/10.1038/npre.2011.5759.1

Rose S, Borchert O, Mitchell S, Connelly S (2020) *Zero trust architecture* (Tech. Rep.). Gaithersburg, MD: National Institute of Standards and Technology (NIST).

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2020) GenBank. Nucleic Acids Res 49(D1):D92–D96. Retrieved from https://doi.org/10.1093/nar/gkaa1023

Scholz AH, Freitag J, Lyal CHC, Sara R, Cepeda ML, Cancio I, ..., Overmann J (2022) Multilateral benefit-sharing from digital sequence information will support both science and biodiversity conservation. Nat Commun 13(1):1086. Retrieved from https://doi.org/10.1038/s41467-022-28594-0

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. In Gene prediction: Methods and protocols (pp. 227–245). New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-1-4939-9173-0 14

Shiraishi F, Perles V, Yassuda H, Kimura L, Andrade E, Simplicio M (2021) Torrente, a micropayment based Bittorrent extension to mitigate free riding. Proc xxi Braz Symp Inf Comput Syst Security (sbseg) (pp. 82–89). Porto Alegre/RS, Brazil: SBC.

Shrivas M (2019) The disruptive blockchain: types, platforms and applications. Texila Int J AcadRes 17–39.

SisGen (2017) SisGen - Brazilian national system for the management of genetic heritage and associated traditional knowledge. https://sisgen.gov.br/. Accessed 31 Mar 2022

Staffas L, Gustavsson M, McCormick K (2013) Strategies and policies for the bioeconomy and bio-based economy: an analysis of official national approaches. Sustainability 5(6), 2751–2769. Retrieved from https://www.mdpi.com/2071-1050/5/6/2751

ter Steege H, Pitman NCA., Sabatier D, Baraloto C, Salom˜ao, RP, Guevara JE, ..., Silman MR (2013) Hyperdominance in the Amazonian tree flora. Science, 342(6156), 1243092. Retrieved from https://doi.org/10.1126/science.1243092

Strand J, Soares-Filho B, Costa MH., Oliveira U, Ribeiro SC, Pires F, ..., Toman M (2018) Spatially explicit valuation of the Brazilian Amazon forest's ecosystem services. Nat Sustain 1(11):657–664. Retrieved from https://doi.org/10.1038/s41893-018-0175-0

Swan M (2015) *Blockchain: blueprint for a new economy*. O'Reilly Media, Inc.

Thiebes S, Kannengießer N, SchmidtKraepelin M, Sunyaev A (2020) Beyond data markets: opportunities and challenges for distributed ledger technology in genomics. Proc 53rd Hawaii Int Conf Syst Sci 3:3275–3284. https://doi.org/10.24251/hicss.2020.400

Toyama D, de Morais MAB, Ramos FC, Zanphorlin LM, Tonoli CCC, Balula AF, ..., Henrique-Silva F (2018) A novel β-glucosidase isolated from the microbial metagenome of Lake Poraquê (Amazon, Brazil). Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1866(4):569–579. Retrieved from https://doi.org/10.1016/j.bbapap.2018.02.001

UNDP (2021) A pilot to improve genetic resources traceability through blockchain technology launched by the UNDP GEF Global ABS project. https://bit.ly/3x2Mphv. Accessed 20 Mar 2023

Wang C, Chu X (2020) Performance characterization and bottleneck analysis of hyperledger fabric. Ieee 40th Int Conf Distributed Comput Syst (icdcs) (p. 1281–1286).

Watsa M, Erkenswick GA, Pomerantz A, Prost S (2020) Portable sequencing as a teaching tool in conservation and biodiversity research. PLoS Biol 18(4):1–9. Retrieved from https://doi.org/10.1371/journal.pbio.3000667

Wolchok S, Halderman A (2010) Crawling BitTorrent DHTs for fun and profit. *4th usenix workshop on offensive technologies.*

Zutshi A, Grilo A, Nodehi T (2021) The value proposition of blockchain technologies and its impact on digital platforms. Comput Ind Eng 155:107187. Retrieved from https://doi.org/10.1016/j.cie.2021.107187