Amazon Biobank: a Blockchain-based Genomic Database for Bioeconomy

Leonardo T. Kimura[1], **Marcos A. Simplício Jr.**[1]

[1]Escola Politécnica - Universidade de São Paulo (USP)

{lkimura, mjunior}@larc.usp.br

Abstract. *The bioeconomy, an industrial production model based on biological resources and sustainable development, can be considered an emerging opportunity for biodiversity-abundant regions, such as the Amazon rainforest. However, existing genomic repositories lack 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; meanwhile, by using smart contracts directly registered in the system, we enforce fair benefit-sharing among all system participants. Moreover, 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 evaluated 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.*

1. Introduction and Motivation

High-biodiversity regions have great potential to develop economic activities in a sustainable manner [Nobre and Nobre 2019, Nobre et al. 2016]. The Amazon rainforest alone, for example, provides critical ecological services whose annual value is estimated to be worth trillion dollars [Strand et al. 2018]. Moreover, the biodiversity in these regions, built over millions of years of evolution [Hoorn et al. 2010], has the potential to stimulate biotechnological development in several fields. Examples include biomimetic engineering, synthetic biology, and the development of new materials, chemical compounds, and biofuels [Nobre and Nobre 2019, Rech 2011].

However, there are still significant challenges to fostering bioeconomic development in high-biodiversity regions. One is the scale of effort necessary to survey millions of species across extensive forest areas – e.g., the Amazon rainforest exceeds 5 million square kilometers. Thus, the collaboration of local residents is of immense value, especially in critical activities such as identifying species with specific properties (e.g., medicinal). Nevertheless, there is not much incentive for such collaboration. After all, it is not always obvious how to ensure fair compensation for these residents. In addition, this collaboration is hindered by the practice of biopiracy in deprived areas [Mgbeoji 2007], as had occurred in the field of genomic research [Li 2021]. Partly for this reason, the principle of benefit-sharing is one of the main objectives of the Convention on Biological Diversity (CDB), a global agreement that aims at the conservation and the sustainable use of biodiversity [Glowka et al. 1994]. This principle can thus be seen as a strong requirement to better promote data sharing and, ultimately, promote biotechnology initiatives.

One proposal in that direction is to build a collaborative and highly scalable genomic database. This database could be populated by any resident of areas of interest, who would retain data ownership and receive appropriate compensation for their contribution. This approach contrasts with (and also complements) the many genomic repositories that currently support biotechnological research. For example, the US National Center for Biotechnology Information (NCBI) maintains a genomic database, which is also done by the European Bioinformatics Institute (EBI) in Europe, and the DNA Databank of Japan (DDBJ). However, these repositories make data publicly available without any kind of usage tracing. This model may facilitate data re-usage, but it does not contribute to an adequate sharing of economic benefits. For example, even if a highly profitable medicine is developed using genomic data from these repositories, their profits are usually not distributed to the corresponding data provider. Consequently, 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.

To address these issues, recent works in the literature have suggested the deployment of collaborative technologies as an integral part of genomic repositories. For example, many studies discuss the potential benefits of blockchain for healthcare genomics [Ozercan et al. 2018, Alghazwi et al. 2022]. In this scenario, blockchain would not simply be an overhyped technology but could be used as a transparent and verifiable record of transactions involving digital assets (e.g., DNA data). Moreover, with the development of a special-purpose currency, blockchain could contribute to fair benefit-sharing among all players of the system. Some of the opportunities created by the technology include data integrity, data ownership (i.e., the owner controls the use of the data), and decentralization (to avoid a single point of failure or to enable distributed data processing).

These characteristics of blockchains motivated us to develop the Amazon Biobank, a community-based genetic database designed to better support biodiversity research in high-biodiversity regions. This would result in (1) larger data variety, since data providers would be compensated for their contribution; and (2) cost reduction, since it is potentially cheaper to purchase genomic data from a database than to organize an expedition to the remote places where that information is found. Amazon Biobank uses blockchain to transparently trace biotechnology products and research to genomic data in the repository. It also uses smart contracts to appropriately share the benefits among all the participants that collect, insert, process, store, and validate genomic data. In addition, it uses other peer-to-peer (P2P) technologies like BitTorrent [Cohen 2003] to build a highly scalable and collaborative computing environment, in which users can contribute (and be paid for) genomic data, computational, and bandwidth resources. Finally, it provides auditability not only for internal system managers but also for any external users. Thus, the correct operation of the system does not critically depend on system administrators, making the architecture adherent to the zero-trust principle.

2. Objective

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.

Therefore, this research seeks to answer the following research question: ***"How to build a genetic database with transparency, scalability, and benefit-sharing properties?"***.

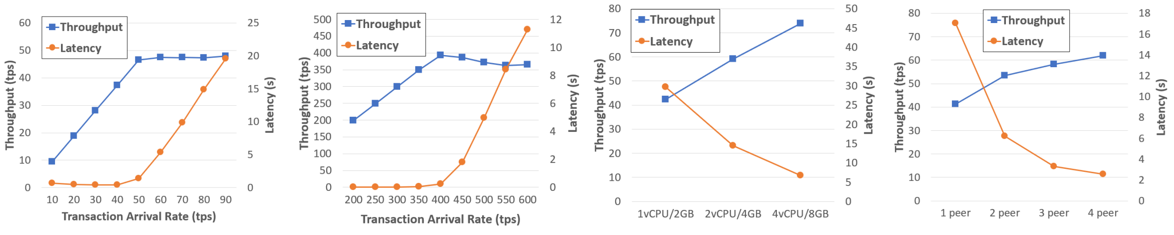
3. Contributions and Results

Our main contribution is the Amazon Biobank proposal, a community-based genetic database supporting biotechnology research. We present its architecture, main modules, and functionalities, and we also describe its main players, operations, and how to avoid some common attacks. By using blockchain and peer-to-peer (P2P) technologies, Amazon Biobank archives the following properties:

- *Benefit-sharing:* By combining blockchain and smart contract, Amazon Biobank implements adequate benefit-sharing among participants who collect, insert, process, store, and validate genomic data. These participants receive biocoins, our internal currency, that can be converted into fiat currencies. The larger the amount of data inserted into the system, the more the value of biocoins, according to supply and demand rules.
- *Traceability:* All operations are recorded in the blockchain to create a temporal and transparent log of events. Thus, products and research can be easily and transparently traced back to the corresponding entry in Biobank. These features are useful, for example, for providing certification of origin, and reproducibility, or for solving disputes involving intellectual property protection.
- *Auditability:* Anyone can audit the blockchain to verify the correctness of the operations registered. Thus, once data entry is registered in the system, its contents or place in time cannot be modified without being detected. This transparency is independent of the amount of trust in the system administration or in its users, resulting in a zero-trust architecture.
- *Scalability:* By using collaborative distributed technologies, like P2P file sharing, Amazon Biobank creates a highly scalable and collaborative computing environment where users can contribute with (and get remunerated for) genomic data and computational, storage, and bandwidth capabilities.

We also assessed the feasibility of Amazon Biobank by building a prototype and evaluating its performance. Our prototype included the main modules involving the blockchain and BitTorrent layer and implements main operations over genetic data, including data registration, purchase, and download[1]. Then, we analyzed its performance (Figure 1). Our experiments showed that Amazon Biobank can handle 400 read and 50 write transactions per second even with a basic configuration. We showed also that its performance can be significantly improved if necessary by increasing the computational resources (vertical scaling) or the number of peers and organizations (horizontal scaling).

[1] The prototype and its documentation are available at https://github.com/amazon-biobank/biobank



(a) Data write **(b)** Data read **(c)** Resources **(d)** Peers

Figure 1. Blockchain prototype performance in: (a,b) base configuration; (c) improving computing resources; (d) increasing the number of peers in each organization.

In addition, we analyzed a hashchain-based micropayment scheme for P2P file sharing [Shiraishi et al. 2021]. We integrated it with Amazon Biobank and evaluated its performance, detecting the overhead of 10% on the download time compared to without micropayments. Our experiments also showed that the addition of one extra seeder reduces the download time by nearly 30%. Therefore, we concluded that the benefits of financially encouraging new seeders are enough to compensate for the micropayment overhead.

Finally, we explain how Amazon Biobank ensures blockchain correctness, allowing independent auditors to detect attempts to manipulate the blockchain. For example, inadequate endorsing attacks, in which some malicious endorsing peers approve an invalid transaction, can be detected by re-executing the transactions stored in the blockchain. As these attacks are not specific to Amazon Biobank, they can be also applied to other permissioned blockchain-based applications.

4. Overview of the Amazon Biobank

The main goal of Amazon Biobank is to promote the collaborative development of biotechnology research in regions with rich ecosystems. Figure 2 describes the main roles played by the entities that compose the system.

Collectors are responsible for one of the main operations of the system, the registration of raw DNA data. Typically, residents in regions of high-biodiversity extract raw DNA data using a portable sequencing device and enter them into the Amazon Biobank application. The application encrypts the genomic data and creates a ".torrent" file with the corresponding magnet link. This magnet link is recorded in the blockchain with any relevant metadata entered by the Collector (e.g., common name, place of extraction). Distributors can then use the magnet link to download (and later upload) the encrypted DNA data via BitTorrent. Moreover, Collectors may allow Curators to access plaintext data and metadata to assess and endorse its correctness, adding value to the corresponding records.

Raw DNA data typically require computational processing (e.g., assembling and sequencing) to be more usable in biotechnology research. Collectors can therefore outsource this task to Processors, players who offer their computing power in exchange for a reward. Processors contact Distributors to download raw DNA data, and then submit the processing results to the system as processed DNA. To avoid any misconduct (e.g., recording data that does not correspond to the raw DNA), these results are verified by the

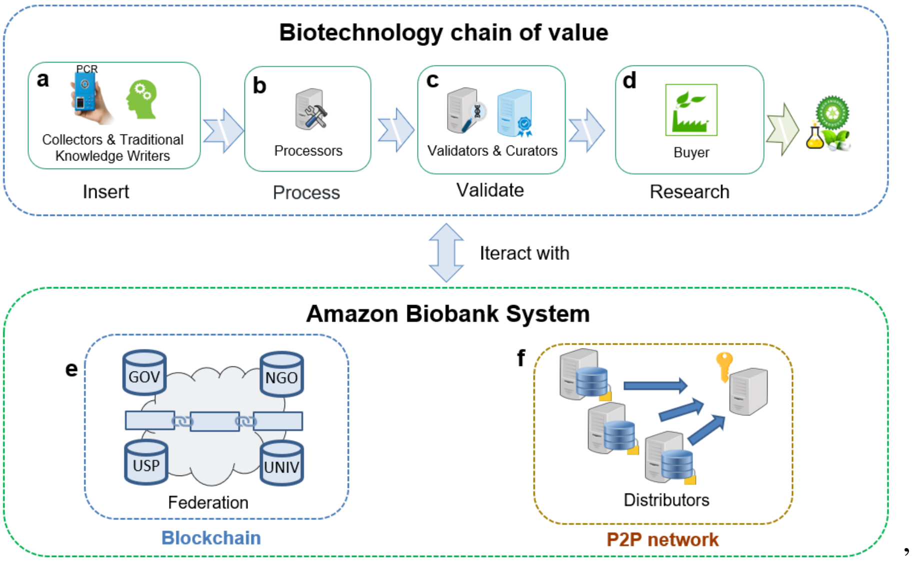


Figure 2. Amazon Biobank: overview and main operations

Validators, which are other Processors who have worked simultaneously or subsequently on the same DNA sequences. Any malicious behavior is punished accordingly, either by suspension of rewards, loss of reputation, or even eviction from the system.

Finally, Buyers gain access to DNA data by paying some biocoins, the Biobank's internal cryptocurrency. This purchase is registered in the blockchain, and the paid biocoins are distributed via smart contracts to all entities involved in the acquisition and processing of DNA data (Collectors, Processors, Validators, and Curators). To help find DNA data of interest, Amazon Biobank supports some searching procedures, based on keywords matching the corresponding metadata, or on genomic sequences matching the DNA data content.

5. Associated Scientific production

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

- **Journal Article:**
  - IEEE Access (2024) *"Amazon Biobank: Assessing the Implementation of a Blockchain-based Genomic Database."* [Kimura et al. 2024]
  - Functional & Integrative Genomics (2023) *"Amazon Biobank: A collaborative genetic database for bioeconomy development."* [Kimura et al. 2023]
- **Conference Papers:**
  - XXI Brazilian Symposium on Information and Computational Systems Security (2021) *"Amazon Biobank - A community-based genetic database."* [Kimura et al. 2021]
  - XXI Brazilian Symposium on Information and Computational Systems Security (2021) *"Torrente, a micropayment-based Bittorrent extension to mitigate free riding"*. [Shiraishi et al. 2021]

The complete master thesis is available in `https://doi.org/10.11606/D.3.2023.tde-26102023-100339`

Table 1. Some biobank properties relevant to biotechnology research

Property	Description
Data validation	Check the correctness of the data or metadata entered
Distributed processing	Outsource the genetical data assembling and sequencing to other users
Sequence search	Query genetic data based on similarity to a sequence of interest
Benefit sharing	Distribute the economic benefits to all involved players
Owner association	Associate each uploaded DNA sequence with its uploader

Table 2. Comparison between blockchain-related genetic projects

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

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

6. Related Works

Many blockchain-based genomic repositories aim to remove brokers and increase user control over their data. Table 2 lists relevant examples, comparing some of their features that are relevant to biodiversity research: support for data validation, distributed processing, sequence search, benefit sharing, and an association between the data and its owner (see Table 1).

For example, one of the analyzed genomic marketplaces, Nebula Genomics [Grishin et al. 2018], is a platform in which users can provide their genetic data in exchange for cryptocurrency tokens. However, Nebula's business model tends to decrease user control over genomic data: once the data is registered in the system, users have little visibility or oversight over how it is handled. In contrast, Amazon Biobank allows users to better control their genomic data through configurable smart contracts, defining the price and the conditions for its use.

Note that many of these biotechnology platforms provide limited support for intellectual property protection or benefit sharing via royalty payments. This is because these platforms prioritize human genetics, rather than focusing on biodiversity as an asset. In addition, 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, in Amazon Biobank, the anonymization of data owners possibly 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 the traceability of genomic resources and benefit-sharing [UNDP 2021]. With the major goal of implementing the Nagoya Protocol [Buck and Hamilton 2011], the Global ABS Tracker project is currently in the early stages, with a pilot prototype launched. The project, nonetheless, tries to handle all kinds of natural products, such as plants or natural substances, and does not focus solely on genetic data. Hence, the system does not support collaborative and private storage of genomic data, nor the analysis, validation, and search of DNA sequences. Also, one of the challenges of the project is that it requires global coordination among 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 trust 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.

The next steps for Amazon Biobank include a broader test evaluation in collaboration with other universities or non-profit organizations (such as the Amazon 4.0 Institute). This deployment will result in a more mature specification that can be used to implement a more definitive, production-ready version of the Amazon Biobank. If successful, the Amazon Biobank will be an important tool for promoting biotechnology research and unlocking the potential of high biodiversity regions, such as the Amazon Rainforest.

References

Alghazwi, M., Turkmen, F., Van Der Velde, J., and Karastoyanova, D. (2022). Blockchain for genomics: A systematic literature review. *Distrib. Ledger Technol.*, 1(2).

Buck, M. and 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.

Cohen, B. (2003). Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, pages 68–72.

Glowka, L., Burhenne-Guilmin, F., Synge, H., McNeely, J. A., and Gündling, L. (1994). A guide to the convention on biological diversity.

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

Hoorn, C., Wesselingh, F. P., ter Steege, H., Bermudez, M. A., Mora, A., Sevink, J., Sanmartín, I., Sanchez-Meseguer, A., Anderson, C. L., Figueiredo, J. P., Jaramillo, C., Riff, D., Negri, F. R., Hooghiemstra, H., Lundberg, J., Stadler, T., Särkinen, T., and Antonelli, A. (2010). Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*, 330(6006):927–931.

Kimura, L., Andrade, E., Carvalho, T., and Junior, M. S. (2021). Amazon biobank - a community-based genetic database. In *Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg)*, pages 74–81, Porto Alegre/RS, Brazil. SBC.

Kimura, L. T., Andrade, E. R., Nobre, I., Nobre, C. A., de Medeiros, B. A. S., Riaño-Pachón, D. M., Shiraishi, F. K., Carvalho, T. C. M. B., and Simplicio, M. A. (2023). Amazon biobank: a collaborative genetic database for bioeconomy development. *Functional & Integrative Genomics*, 23(2):101.

Kimura, L. T., Shiraishi, F. K., Andrade, E. R., Carvalho, T. C. M. B., and Simplicio, M. A. (2024). Amazon biobank: Assessing the implementation of a blockchain-based genomic database. *IEEE Access*, 12:9632–9647.

Li, F.-W. (2021). Decolonizing botanical genomics. *Nature Plants*, 7(12):1542–1543.

Mgbeoji, I. (2007). *Global biopiracy: patents, plants, and indigenous knowledge*. ubc Press.

Nobre, C. A., Sampaio, G., Borma, L. S., Castilla-Rubio, J. C., Silva, J. S., and Cardoso, M. (2016). 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*, 113(39):10759–10768.

Nobre, I. and Nobre, C. A. (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*.

Ozercan, H. I., Ileri, A. M., Ayday, E., and Alkan, C. (2018). Realizing the potential of blockchain technologies in genomics. *Genome research*, 28(9):1255–1263.

Rech, E. (2011). Genomics and synthetic biology as a viable option to intensify sustainable use of biodiversity. *Nature Precedings*.

Shiraishi, F., Perles, V., Yassuda, H., Kimura, L., Andrade, E., and Simplicio, M. (2021). Torrente, a micropayment based Bittorrent extension to mitigate free riding. In *Proc. of the XXI Brazilian Symposium on Information and Computational Systems Security (SBSeg)*, pages 82–89, Porto Alegre/RS, Brazil. SBC.

Strand, J., Soares-Filho, B., Costa, M. H., Oliveira, U., Ribeiro, S. C., Pires, G. F., Oliveira, A., Rajão, R., May, P., van der Hoff, R., Siikamäki, J., da Motta, R. S., and Toman, M. (2018). Spatially explicit valuation of the Brazilian Amazon forest's ecosystem services. *Nature Sustainability*, 1(11):657–664.

UNDP (2021). A pilot to improve genetic resources traceability through blockchain technology launched by the UNDP GEF Global ABS project. `https://bit.ly/3hnMqEh`. Acessed on 29-06-2021.