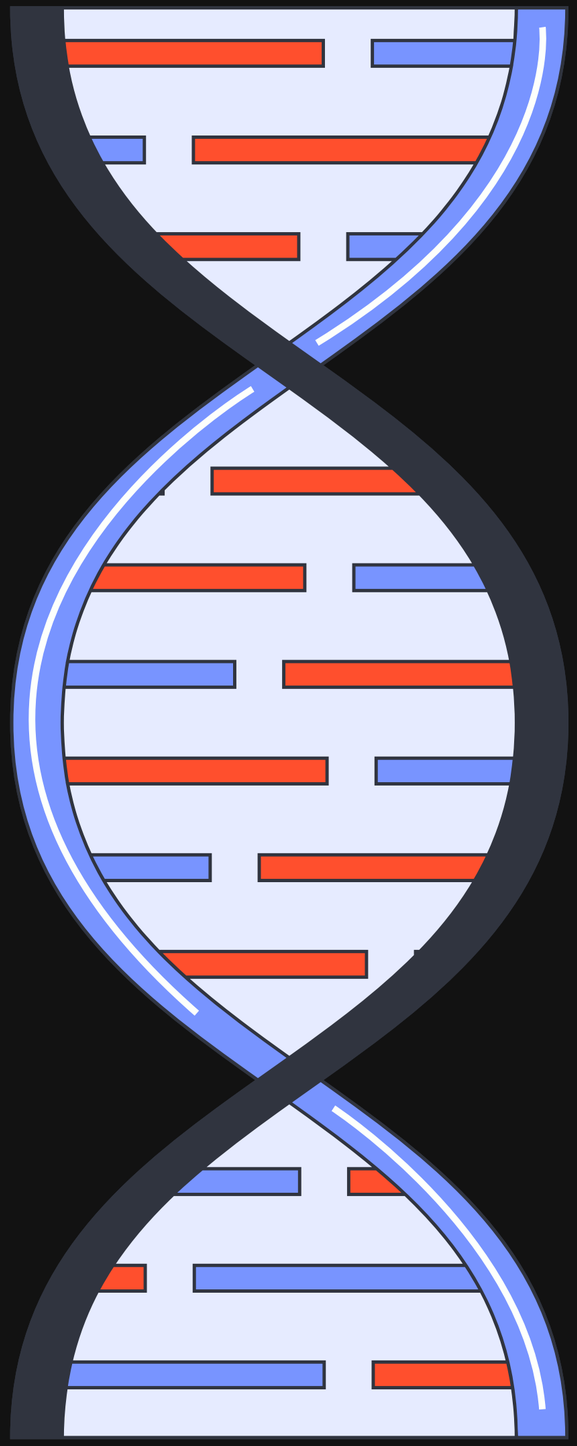
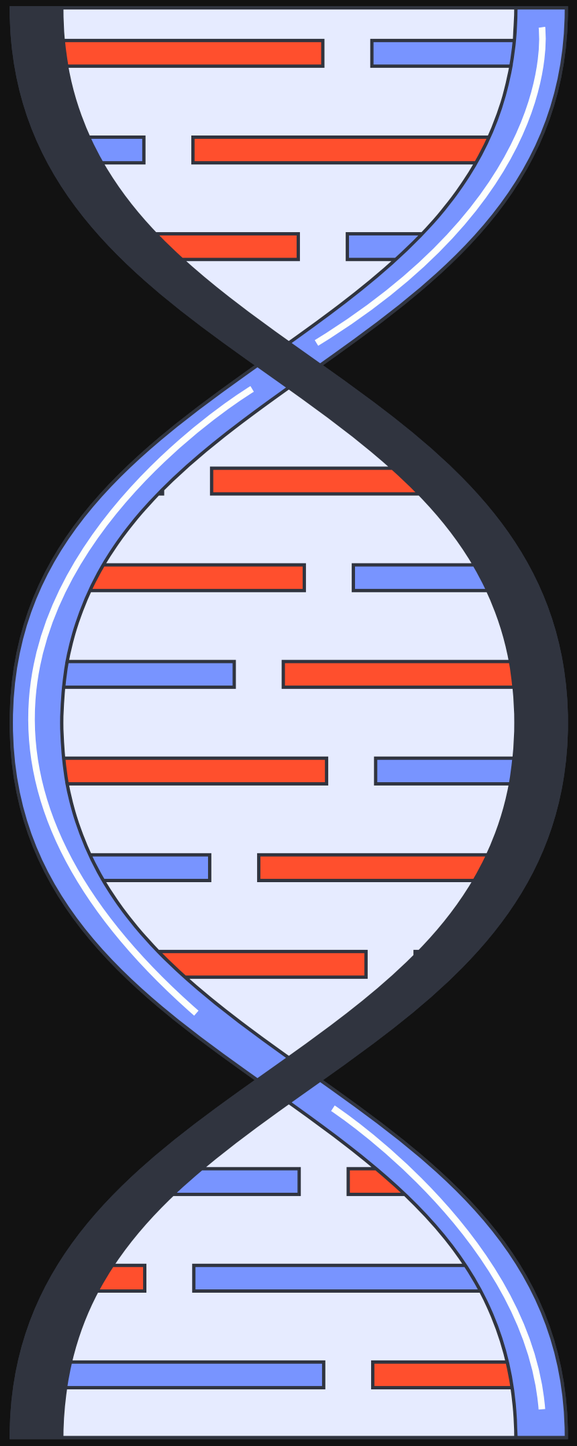


FINAL EVALUATION

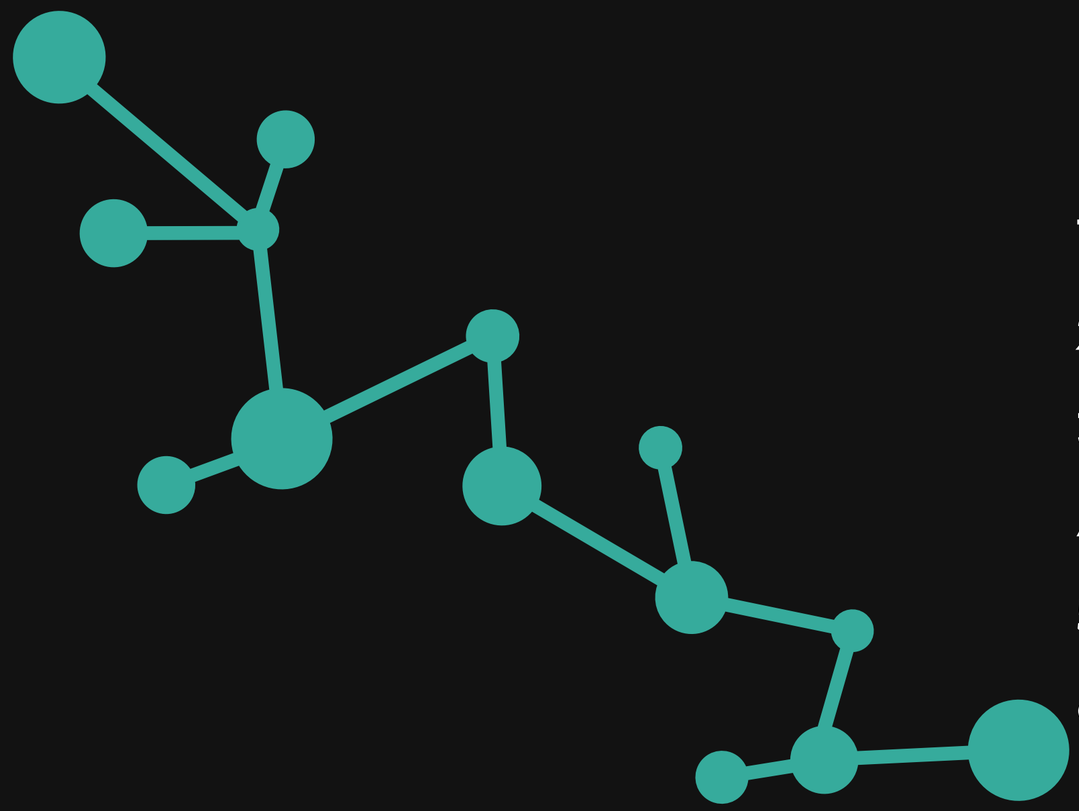
PRACTICAL BIOINFORMATICS

Group 11

GENOMIC DATA PRIVACY



PRESENTATION OUTLINE



1: PROBLEM

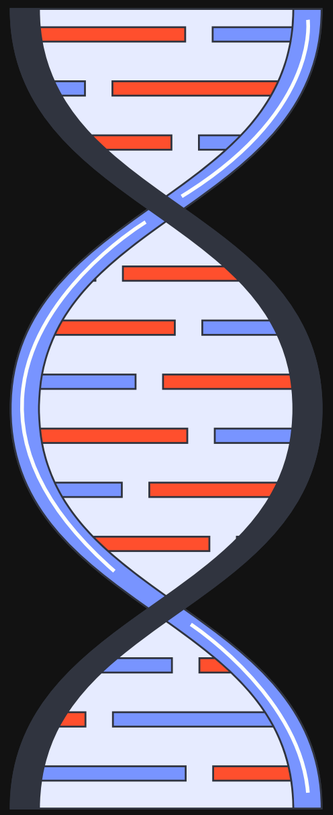
2: SOLUTION AIM

3: APPLICATION OF SOLUTION

4: CITATIONS

5: DEMO

6: Q/A SESSION



PROBLEM

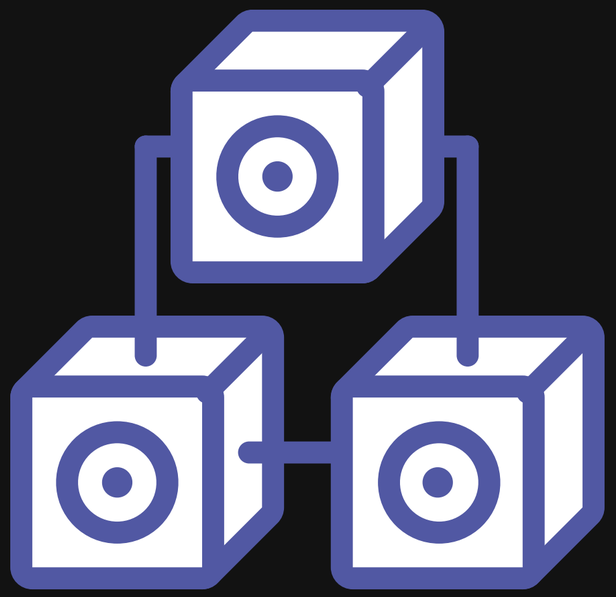
Current databases are maintained and owned by 1 company that can sell or misuse our data. In essence, the current data is centralised. If there's ever any loss of data then that would lead to all data being lost with no possible backups.

This data can easily be exploited by companies, advertisers, hackers and government for exploiting people for their own needs.

Users opt in to these business models (such as 23AndMe, Ancestry etc.) without knowing the consequences of sharing this data freely.

Users don't have effective control over their genomic data. It is owned by multi national corporations and leaves great possibility of the said data being misused.

SOLUTION AIM



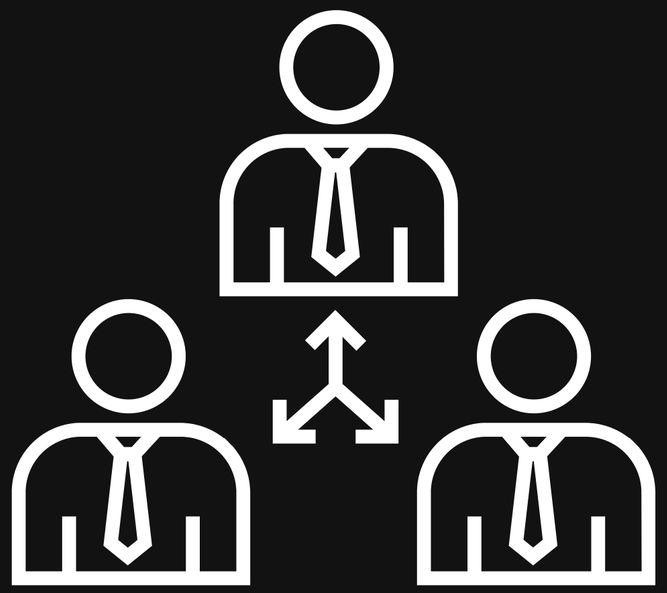
Our solution to the problem aims to do the following:

1. Attempt to give effective control of ones genomic data to the person themselves.
2. Attempt to expand biological research sample collection with our model.
3. Attemp to stop proliferation of data by companies for profit.
4. Maintain a permanent repository of all biological transactions.
5. Attempt to gradually create a system of profits and incentives via genetic data submission.

STATUS QUO

No Privacy

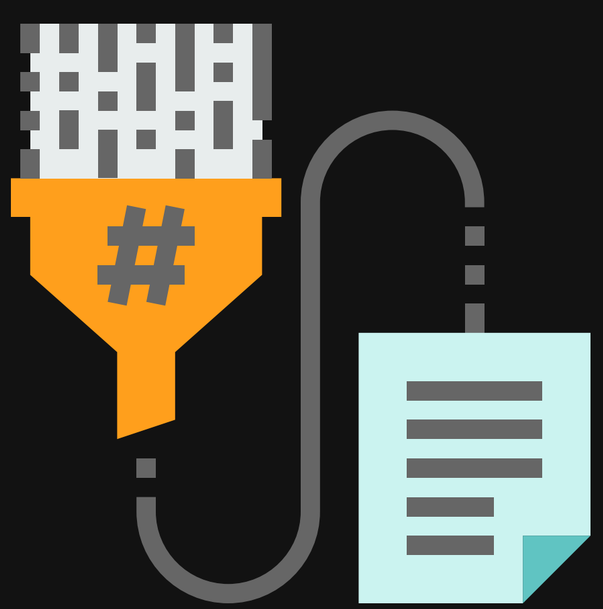
BLOCKCHAIN



Where exactly does blockchain come into the picture?

1. This is done to make sure that the data now being transferred is kept in a record permanently although through a chain of hashes and keys existing on the blockchain
2. A record of any valid transaction will store the hash of the uploaded encrypted file on the IPFS, the sender and the receiver's public key.
3. Now data can never truly be lost since the item is now stored in a group of computers. So even if one computer fails the nodes will still be alive.

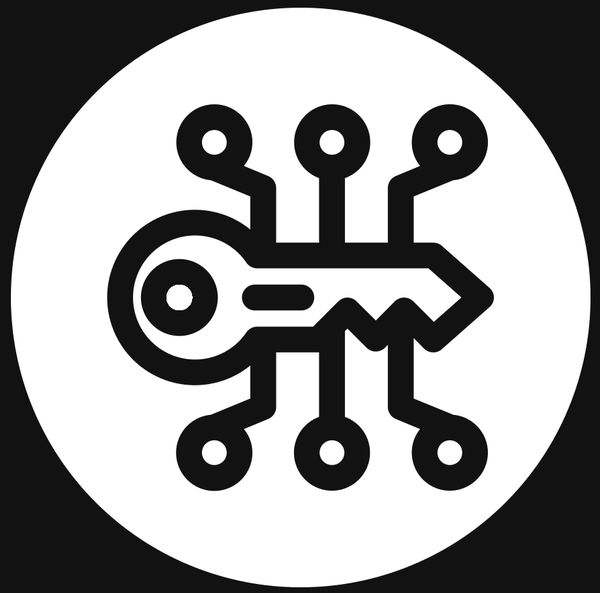
IPFS



Where does IPFS come in?

1. Currently a drawback of blockchain is that it cannot store large chunks of data, due to its processing speed. So how to store a comparatively large chunk of data?
2. Enter IPFS. It's a peer-to-peer protocol where each node stores a collection of hashed files. A client who wants to retrieve any of those files enjoys access to a nice abstraction layer where it simply needs to call the hash of the file it wants. IPFS then combs through the nodes and supplies the client with the file.

GPG: ASYMETRIC ENCRYPTION

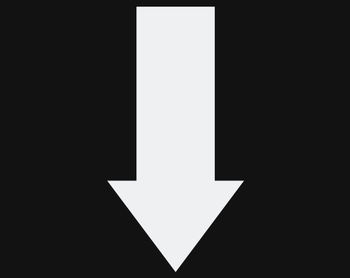


Why encryption?
One of the big issues with blockchain and IPFS is that everything is publically available and this can be exploited if anyone can view our genetic data.

But, what if we have a mechanism wherein we create an encrypted file that can only be accessed by a specific set of people whom we want.

This is where asymetric encryption comes in. And we will be using the GNU Privacy Guard to show how this concept works.

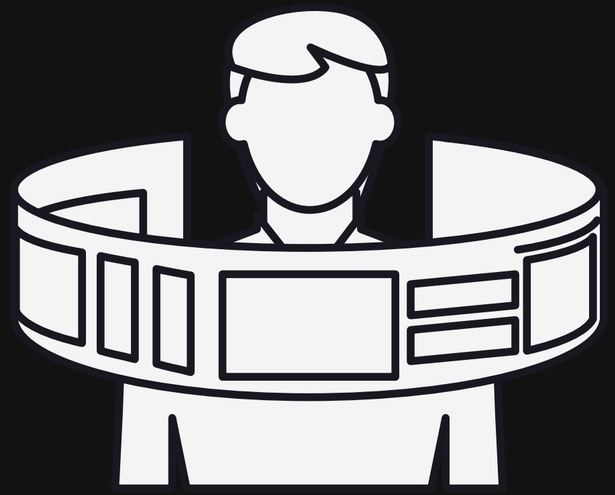
OUR MODEL



Organisation will have access to the key

Privacy
Ownership

EFFECTIVE CONTROL



We now shift the control over one's genomic data from companies such as 23&me to the individual themselves. They will be the ones uploading their genetic data on the IPFS and they will be the ones encrypting it and they will be the ones giving access to the companies now.

NO ILLEGAL PROLIFERATION OF DATA

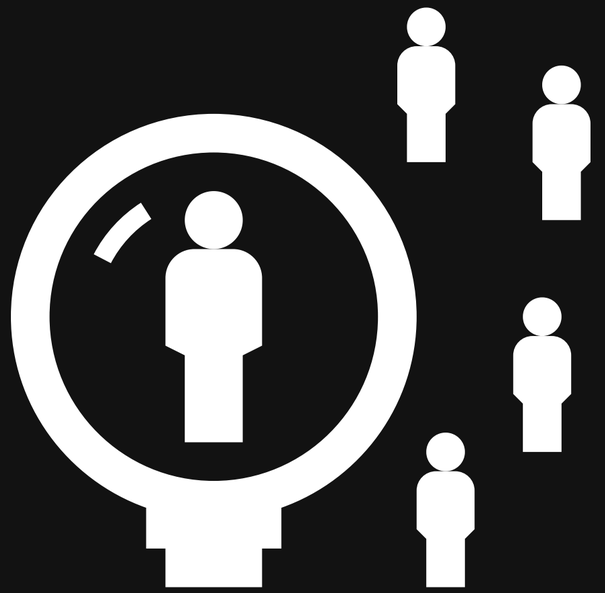
This will be possible simply because for any company to hypothetically share all their data with some organisation or company they will effectively have to give away their private key to the other alias and only then can the data be considered to be sold.

However get that giving away a private key is essentially the same as giving away the pin to your credit card. Hence the tendency for a company sharing their private key and GPG creds is very less likely.

PERMANENCY OF DATA

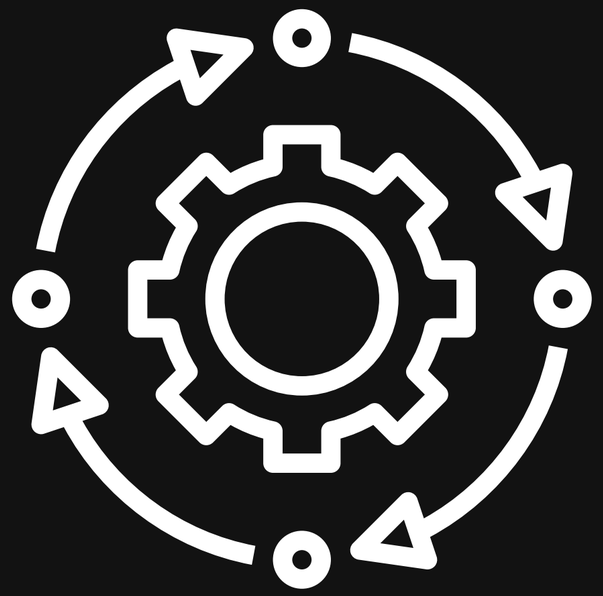
Storing the hash of our file uploaded on the IPFS not only makes sure that the data is not under the control of one particular person but also makes sure at the same time that the data is never removed since even if one nodes crashes from a group of nodes, the other nodes keep it alive. Transactions being kept permanently on the blockchain ensures accountability and provides for making an incentive based profit ecosystem.

MORE SAMPLING IN RESEARCH



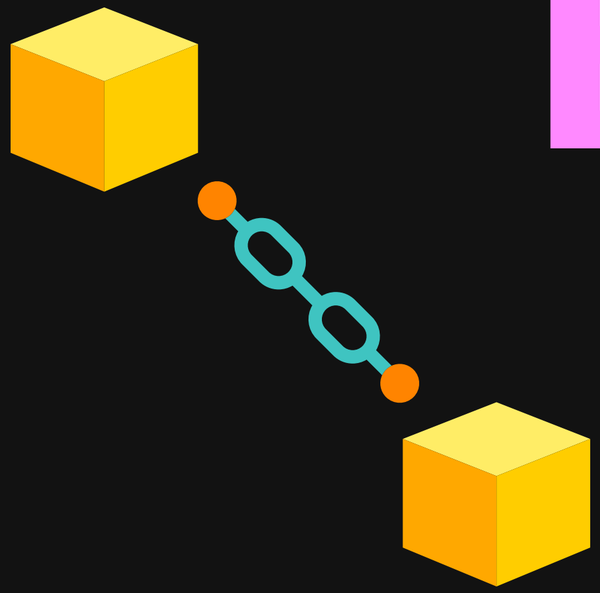
We believe with the security and concreteness that comes with our solution we also make sure that more people will gradually opt in for these biological services and participate in bioinformatics research. We can now cross borders and achievemuch larger human datasets for research purposes.

ABILITY TO CREATE A SYSTEM



We believe if this system is scaled it will gradually be able to create a system of participatory research wherein a person can be paid for their contributions to participating in a research and as a result organisations conducting can get much more larger of a sample size for their papers and experiments.

LITERATURE BACKING



Genesy: a Blockchain-based Platform for DNA Sequencing

Roberto Carlini[1], Federico Carlini[1], Stefano Dalla Palma[2], and Remo Pareschi[3]

[1] Genesy Project, Ferrara, Emilia Romagna 44121, Italy
info@genesyproject.com
[2] Jheronimus Academy of Data Science (JADS), Sint Janssingel 92 5211 DA 's-Hertogenbosch, The Netherlands
s.dallapalma@uvt.nl
[3] University of Molise, C.da Fonte Lappone 86090 Pesche (IS), Italy
remo.pareschi@unimol.it

2 Genesy Ecosystem: How Does it Work?

The aim of Genesy is to involve collaboration among users and various organizations to promote a high-level genomic ecosystem, thereby efficiently collecting and managing the large volumes of data produced in sequencing activities. To this aim the following components have been proposed.

- **Kit for DNA collection.** It consists of a small, light and thin vial that will preserve dry saliva intact for months with reduced logistics costs. It will be delivered to the customer in a simple personalized packaging with our brand. The vials will be regularly insured and sent by batch to the sequencing centers, containing even more shipping costs. Users of the platform will receive their results within a few weeks. A service will be activated that will regularly notify the user of the possibility to acquire new reports applied to his DNA, which will be produced on a par with the development of new genetic panels and related medical research progress.
- **Sequencing structures.** Genesy will sequence DNA at various levels (complete genome and genetic panels) in Italian and US laboratories through medium-high-end machines, therefore for areas such as nutrition, allergies, fitness, microbiome, etc.; and for diagnostic panels, for example autoimmune diseases, physical and neuropsychilogical traits.
- **IBM technology infrastructure.** The IBM Hyeperledger blockchain platform[1] for managing user meta-data is in turn integrated with the following services:
  - *NoSQL database* - to analyze and manage sequenced DNA data;
  - *IBM Watson* - as a tool to create reports on data produced by sequencing;
  - *Cloud Object Storage* - to store data and some off-genomic data;
  - *Network Stellar* - to manage interchange transactions through ad-hoc cryptocurrency.

Design Principle 3 - ecosystem cryptocurrency. A new currency that can be defined as a token utility will allow to purchase and sale data on the Stellar network[2]. This will allow the Genesy company to monetize the sale of its services, activating an exchange market and increasing its value depending on the increase of activities by the whole ecosystem. This will help solve one of the most critical aspects of the genetic data market, that is, how to standardize information and value.

All transactions on the platform will be carried out in Genesy. The Genesy coin can be "exchanged" in different ways and for different reasons. There are two different types of advantages: (1) at the time the DNA is sequenced, some Genesy coins will be credited to the account associated with the user profile; and (2) if the user will share its DNA data on the platform, s/he will be rewarded with Genesy based on the volume of analysis done by pharmaceutical companies and research. As a result, the longer the DNA data will share, the higher the profit in Genesy will be. These analyses will be measured by internal meters, which will distribute Genesy coins among the accounts of all users who will share their DNA. This means that even if the DNA in particular will not be analyzed, it will still earn.

In addition to the immediate compensation for the sharing of DNA and the future variable

2.2 Architecture Elements

The proposed platform provides that genome owners maintain ownership of sequenced data by exploiting the security and immutability features offered by blockchain technology, and that they are basically accessible.

The software architecture of the proposed solution entails a Genome-as-a-service architectural style wherefore a genome data marketed through the proposed ecosystem, that is, for a sequencing service (e.g., to gather information on the risk of a person to contract a specific disease), it is addressed through the marketplace itself, and worked out by peers (i.e., universities research centers, private labs, hospitals, geneticists, pharmaceutical companies, etc.), but, at the same time, the genome data is also sold through the same marketplace. Therefore, whenever someone buys genome data, s/he buys it as "a service", meaning that s/he gets that information (i.e., the genome data) following the pay-per-use schema, while the ownership of the data still belongs to its original producer.

The aforementioned architectural style requires the architecture elements listed below.

- **Blockchain node** - is any node that contain public information about genomimc data and users operating on the platform (e.g., humans, hospitals, research centers, etc.). A blockchain node also contains information about users transactions, for example when they exchange coins or share information through smart contracts.
- **Ecosystem user** - customers, scholars, geneticists, pharmaceutical companies, private labs etc.

HealChain: A Decentralized Data Management System for Mobile Healthcare Using Consortium Blockchain

Weiquan Ni, Xumin Huang, Junxing Zhang, Rong Yu*
Guangdong University of Technology, China
E-mail: weiquanni17@163.com, huangxu_min@163.com, junxingzhang@mail2.gdut.edu.cn, yurong@ieee.org

2.3 Storage Layer

Since blockchain is not designed for large-scale data storage, it is reasonable to store healthcare data into a distributed off-chain storage system. So authentication information and links of healthcare data in the off-chain storage system are only stored on the blockchain.

In this paper, we introduce IPFS as the off-chain storage system for HealChain. Here, IPFS is regarded as a peer-to-peer data storage system that can be designed well to be friendly to the proposed consortium blockchain. IPFS is a distributed file system wherein a mass of computing devices connect by an identical system of files [11]. For secure healthcare data storage, IPFS provides a high-throughput external storage system with mature storage strategies. Healthcare data in IPFS is addressed by its cryptographic hash entry, through which users can conveniently access and download individual healthcare data in HealChain. Due to the significant advantages of IPFS, the links for retrieving the healthcare data always stay the same regardless of which IPFS nodes serve the data. Moreover, the original data can be recovered in time by erasure codes although data stored in IPFS is attacked or destoryed [12].

3.2 Advantages of the Proposed HealChain

We offer necessary security analysis to exhibit great advantages of HealChain. According to the descriptions in [13], some important security and privacy requirements are exactly satisfied in our proposed HealChain. We summarize the great advantages into the following aspects.

- *Confidentiality*: Standard cryptographic primitives including the asymmetric encryption and digital signature are exploited for protecting healthcare data. Thus, the encrypted data cannot be decrypted by any entities without the private keys of users or receivers.
- *Integrity and authentication*: Digital signature also can realize the verification regarding the integrity and ownership of healthcare data. And the digital signature can only be signed by the private key of the users, which guarantees the effectiveness of the digital signature.
- *Availability*: All the CBNs cooperate to audit all the uploaded healthcare data from legal users, which avoids unauthorized access from illegal users and eliminates potential security threats. Moreover, due to distributed network topology of HealChain, users have the right to choose different CBNs for uploading healthcare data toward removing denial of service attack.
- *Traceability*: Owing to blockchain-based nature, all the requests and operations on healthcare data are accountable and traceable in HealChain. Once invalid access, identity fraud and data leakage are detected, the malicious entities can be identified and recorded in time.

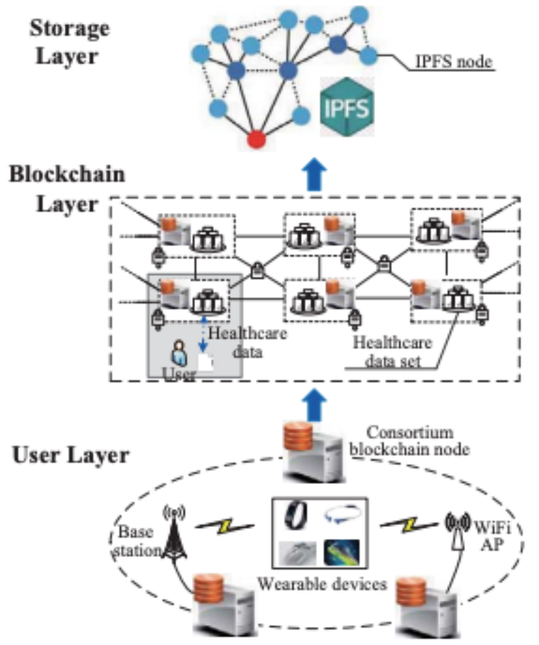
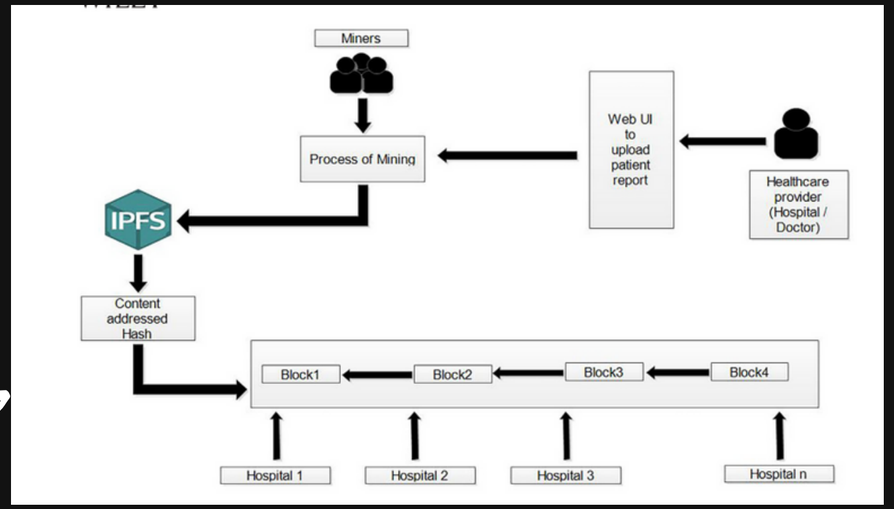


Fig. 1: The architecture of HealChain in mobile healthcare.

Decentralized secure storage of medical records using Blockchain and IPFS: A comparative analysis with future directions

Shivansh Kumar | Aman Kumar Bharti | Ruhul Amin

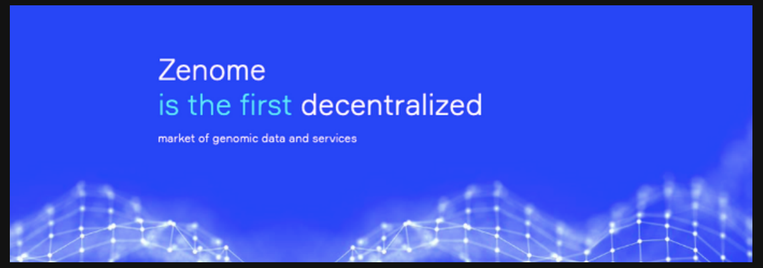


Blockchain for Genomics: A Systematic Literature Review

MOHAMMED ALGHAZWI, University of Groningen, Netherlands
FATIH TURKMEN, University of Groningen, Netherlands
JOERI VAN DER VELDE, University Medical Center Groningen, Netherlands
DIMKA KARASTOYANOVA, University of Groningen, Netherlands

Table 3. Blockchain-based Commercial Genomic Marketplaces

Company	Blockchain Platform	Incentive	Services
Genecoin [47]	Bitcoin blockchain	N/A	backup encrypted DNA to Bitcoin blockchain.
Encrypgen [41]	GeneChain to facilitate cryptocurrency exchange and record data exchange	Cryptocurrency: DNA token	HIPAA-compliant storage of genomic data and exchange platform after de-identifying the data
Embleema [40]	Private Ethereum blockchain to register patient consent and manage data flow	Cryptocurrency: RWE token	Patient consent management, optimize clinical trial recruitment, secure HIPAA-compliant storage, analyze, and organize the data for researchers.
Nebula Genomics [35], [53], [54]	Exonum [43] permissioned blockchain for decentralized access control logs, data storage in third-party storage system	Nebula credits which can be used to purchase health-related information about the genome sequence	User controlled access to genomic data with blockchain log, secure storage and analysis in a blackbox environment.
LunaDNA [80]	Unspecified	Company shares and earnings when data are used	HIPAA-compliant storage, user ownership and control of data, contributions to research and medical research.
Shivom [130]	blockchain agnostic based on combination of blockchain implementations	Cryptocurrency: OMX tokens	Pay-per-use marketplace for genomic data, bioinformatics platform with analysis pipelines for researchers.
Zenome [95]	Ethereum-based blockchain	Cryptocurrency: ZNA token	Secure storage, selling access to genetic data, and buying genetic services.
Genomes.io [50]	Ethereum-based Blockchain	Cryptocurrency: GENE token	Secure personal DNA vault, and financial benefits from contributing anonymously to medical research.
DNAtix [37]	Ethereum-based blockchain (proof-of-concept)	Cryptocurrency: internal ERC-20 based token	Anonymous genetic vault service, anonymous testing and reports on genomic data, and connect anonymously with people with similar genetic traits.
Genesy [20]	Private blockchain based on HyperLedger Fabric	Payment in fiat and cryptocurrency through Stellar [116] and Stripe [119]	Sequencing services, selling access to genomic data, and a blockchain-based ecosystem for the sharing of genomic data.
GenoBank [127]	Ethereum-based blockchain with non-fungible tokens (NFT)	Cryptocurrency: ERC-20 token	Control over genomic data with DNA crypto wallet, and secure platform to process the data.
Longenesis [79]	Bitfury's Exonum blockchain	N/A	Patient consent platform targeted to medical institutions, HIPAA and GDPR compliance data storage,
LifeCode.Ai [63]	Quorum platform [28]	Cryptocurrency: ERC-20 token	Individual health data ownership, trading mechanism for data exchange, and secure decentralized storage and management of health data



23andMe

Vision

Zenome plans to build a decentralized storage system for genomic data provided by network participants and supported financially with the help of internal cryptocurrency. Further, we will start the function of free exchange of genomic and personal data within the network. It will be possible to find a person who has specific (interesting to you) eye color, age, weight, nationality and access to his genomic information (to the non-sensitive part of the genome). Next, we implement a system of questionnaires plus rating system to assess if users haven't provided fake data. And, finally, at one of the later stages we plan to attract large companies and scientific centers interested in buying genomic data, as well as buying access to the storage of Zenome. Thus, in the long term, we want to create the best infrastructure project for the genomic Internet - a kind of "genomic Google", owned by the entire community of Zenome.

What service do you provide?

Due to applicable regulations, 23andMe only offers an Ancestry + Traits Personal Genetic Service in international markets and health reports are not available. The service is in English only. It requires submitting a saliva sample using our saliva collection kit that you send to the lab for analysis.

Our Ancestry + Traits Service helps you understand who you are, where your DNA comes from and your family story. We analyze, compile and distill your DNA information into reports on your Ancestry Composition, Ancestry Detail Reports, Traits, Haplogroups, and Neanderthal Ancestry. We also provide a DNA Relatives tool to enable you to connect with relatives who share similar DNA, and an automatic Family Tree builder.

Mission

We believe that the "era of data" has come. At present, the technologies of artificial intelligence in various fields of activity have developed tremendously. The use of artificial intelligence can remove a person from a huge number of processes, make the work more accurate and minimize the number of errors associated with the human factor. Including, in the near future, artificial intelligence will be the main driver for the development of medicine, diagnostics, the creation of new drugs and personalized products. In this case, for the machine mearning and the use of artificial intelligence, a large amount of correctly structured information is needed and the availability of such information in a concentrated form becomes as valuable as gold once was. That is why we want to create a platform that would accumulate this value ("new gold") and at the same time access to this gold will be provided for all the users of the platform, that prevents monopolization.

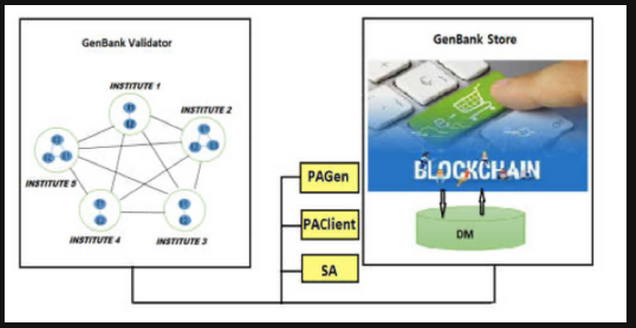
A Blockchain-Based Model of GenBank Store System

Irina Krasteva, Todorka Glushkova
Faculty of Mathematics and Informatics
Plovdiv University "Paisii Hilendarski"
Plovdiv, Bulgaria
glushkova@uni-plovdiv.bg

Nevena Moraliyska
Department of Intelligent Systems
Institute of Information and Communication Technologies,
Sofia, Bulgaria
nevena.uzunova@gmail.com

Nikolaya Velcheva
Information and Documentation Center
Department of Plant Genetic Resources
Institute of Plant Genetic Resources "K. Malkov"
Sadovo, Plovdiv district, Bulgaria
nikolaya_velcheva@abv.bg

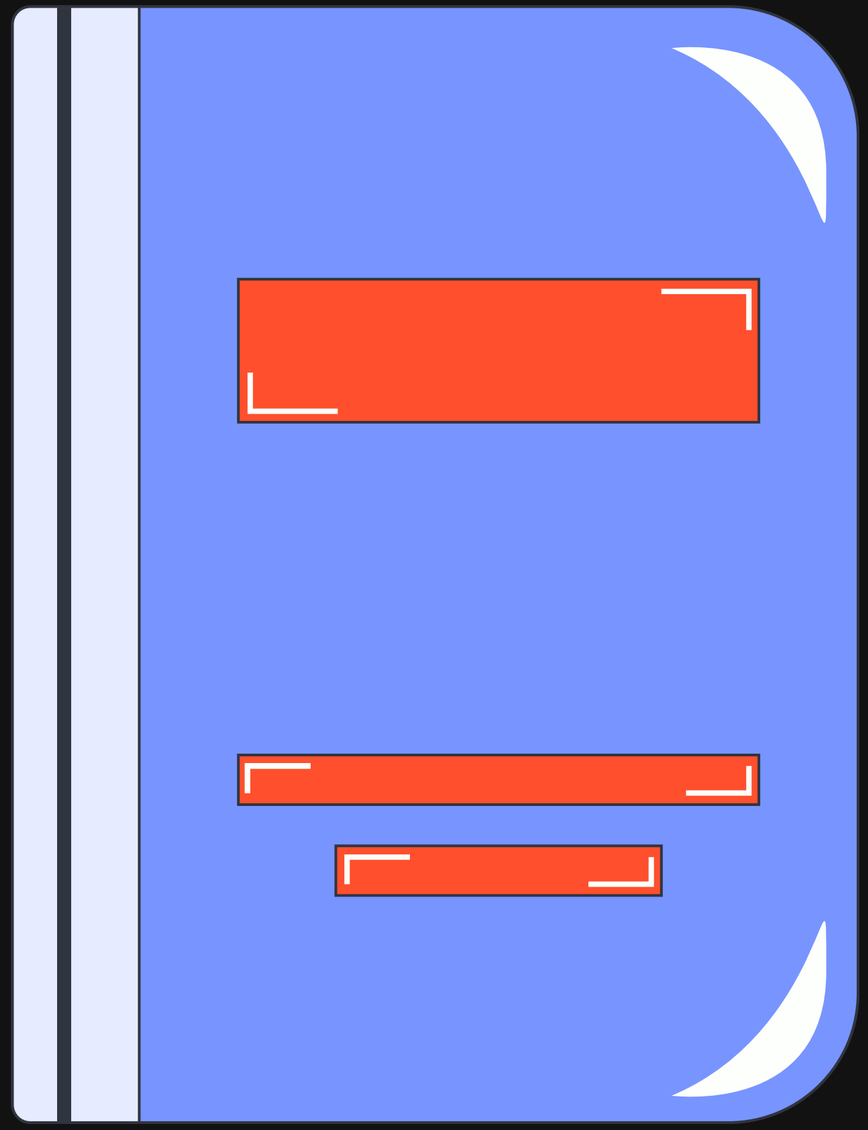
IEEE Research Paper



Zenome Company's Research

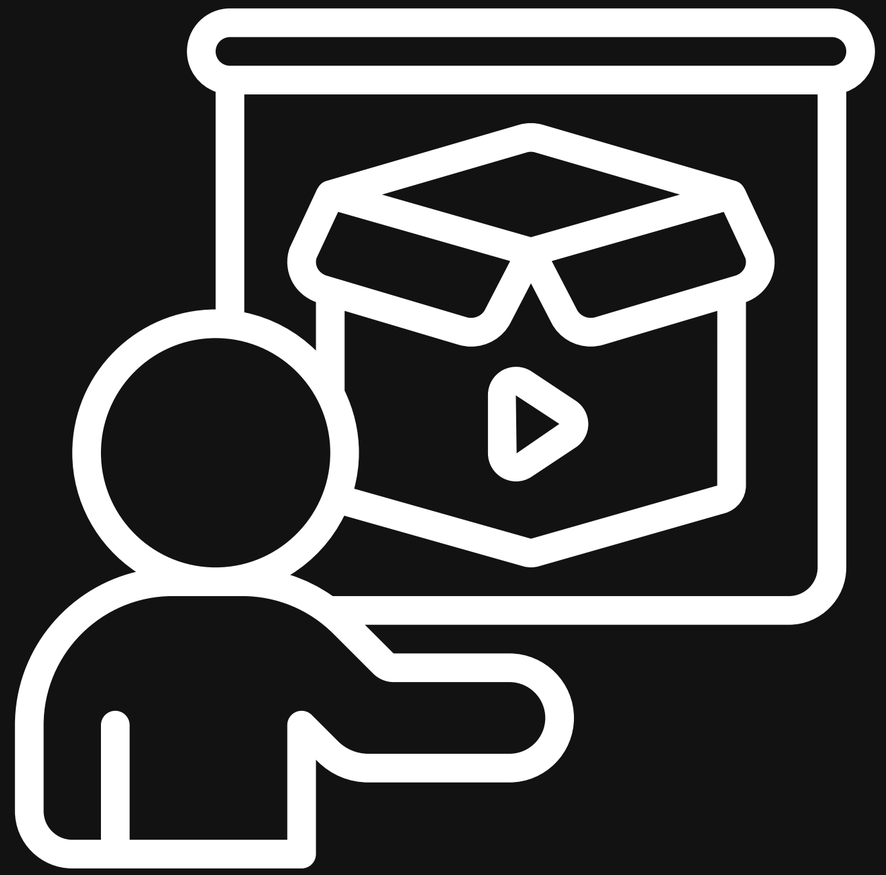
CURRENT PROGRESS

- We have figured out a way that the entire solution can be woven together using encryption and IPFS and blockchain.
- However, currently to make this project scaleable and deployable one would need to create an entire tool which allows for the encryption using GPG through website and automates the entire process. However since GPG is mostly command line based and we are currently hosting our project locally it is not feasible.
- We have created a barebones website where we show how these transactions could be reflected on the blockchain and also done a few exchanges of documents among us to figure out if the encryption is actually effective or not and we will proceed to show this in the Demo.



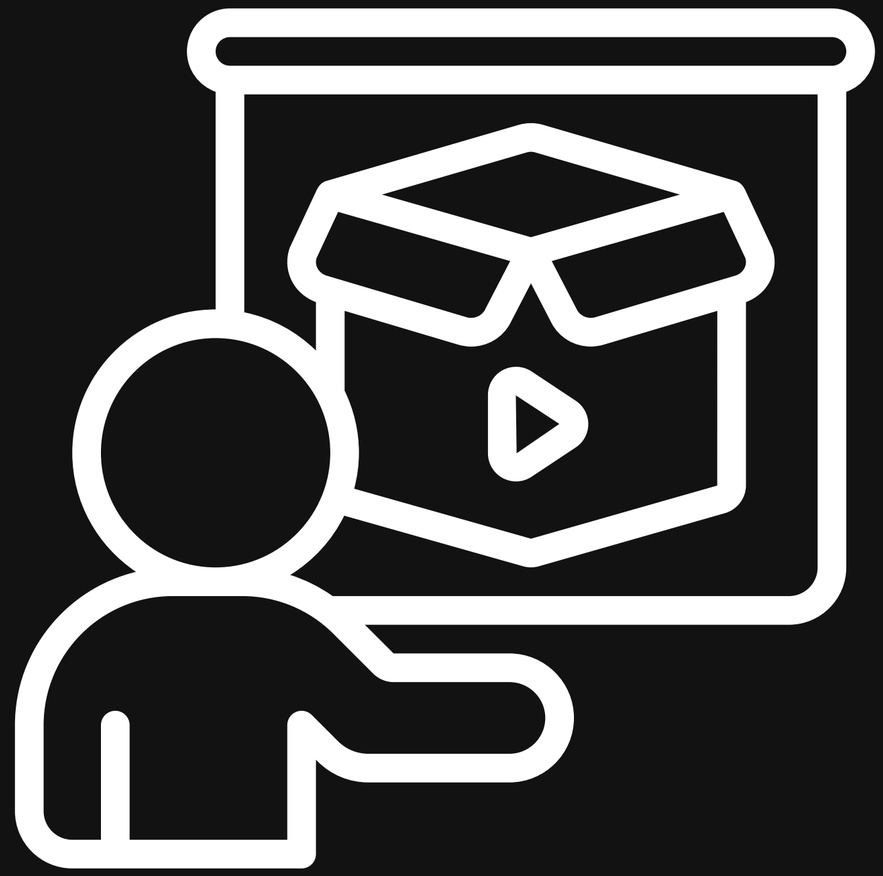
DEMO TIME

- We can now provide a 5 minute demo using multiple computers which would include a sender, a requester and a third party.



CONTRIBUTION

- Abhik : Ideation and Project Flow implementation
- Aishwary: Provided Samples for encryption
- Nikita: Contributed to presentation
- Ojus: Helped in implementation of workflow of project and in presentation
- Vaibhav: Literary research on past work done
- Vishesh: Created the blockchain and implemented the workflow of project



THANK YOU

REFERENCES

1. I. Krasteva, T. Glushkova, N. Moraliyska and N. Velcheva, "A Blockchain-Based Model of GenBank Store System," 2020 IEEE 10th International Conference on Intelligent Systems (IS), 2020, pp. 606-611, doi: 10.1109/IS48319.2020.9200133. (https://ieeexplore.ieee.org/abstract/document/9200133)
2. Zhengxin Chen, Re-examination of a Bioinformatics Database Course: Engaging Blockchain Technology, Procedia Computer Science, Volume 162, 2019, Pages 368-374, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.297.(https://www.sciencedirect.com/science/article/pii/S1877050919320095)
3. Ni, Weiquan, Xumin Huang, Junxing Zhang and Rong Yu. "HealChain: A Decentralized Data Management System for Mobile Healthcare Using Consortium Blockchain." 2019 Chinese Control Conference (CCC) (2019): 6333-6338.
4. Decentralized secure storage of medical records using Blockchain and IPFS: A comparative analysis with future directions
5. http://ceur-ws.org/Vol-2334/DLTpaper6.pdf
6. https://www.genesyproject.com/
7. https://www.23andme.com/en-int
8. https://zenome.io/about/