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

Abstract

Advances in technology have drastically reduced costs and implementation time for Whole Genome Sequencing (WGS). Along with easy access to genomic data, WGS technology can significantly improve the productivity and efficiency of health care and social well-being as well as improving the quality of life and increasing the possibility that people have a direct impact on their health. This paper proposes *Genesy*, an innovative blockchain platform which acts as an intermediary between the owner of the genomic data and its potential users to structure a new ecosystem that incentivize people to share their genomic data, leveraging the potentiality of blockchain technology to safekeep the users' exclusive property and access to their genomic data, allowing them to partecipate in the benefits and the advances of the genomic research.

1 Introduction

Genomics is an area within genetics that concerns the sequencing and analysis of an organism's genome, that is, the whole set of genetic material of an organism. The human genome consists of about 3 billions pairs of DeoxyriboNucleic Acid (DNA) bases distributed on 23 pairs of chromosomes. However, less than 1% of our genome represents the differences between human beings, which can hide several concerns such as tricky-to-identify diseases or disorders. Indeed, an individual's genome contains millions genetic variants that make each person unique. Some contribute to differences in the appearance such as eye color or blood type, whereas some variants have been linked to specific diseases in terms of predisposition and many of them have unknown effects. These variants are detected by comparing the DNA sequences of an individual with a reference genome managed by an international organization such as Genome Reference Consortium. Through the use of these and other reference standards (e.g., American College of Medical Genetics and Genomics and ClinVar - NCBI) it is possible to link the variants to diseases or disorders and to define their severity levels.

Once genomic data is extracted from an individual (e.g., from her saliva), scientists break strands of DNA into individual pieces so that each piece is sequenced individually and reassembled into billions of letters that make up her genome. Advances in DNA sequencing technology have dramatically reduced the costs and time needed to sequence whole exomes and genomes. An exome is the set of all the portions of the genome that "code" for proteins: the DNA present in the exome is the "map" that the cell uses to correctly produce proteins that are fundamental for the cell structure and for the performance of its functions. Although it represents just over 1% of the entire genetic heritage, it is in the exome area and its approximately 20,000 genes that occur over 85% of the mutations known today as clinically relevant. The analysis of the whole exome or part of it (panels) is a very powerful diagnostic tool able to provide

information of enormous general and clinical interest. Genotyping is the most economical DNA sequencing service: genes are compared to a standard reference genome of 500,000 to a million different points to identify the variants. It can explain why an individual have certain somatic traits, shed light on her ancestors as well as reveal some medical risks. Exome sequencing is considerably more detailed and can give important medical scientific information.

Reduced costs and time along with the rise of new digital technologies are profoundly changing the way we can manage knowledge in health-care and are allowing us to have data and the possibility to link data and knowledge that have so far been impracticable. At the same time, they are also raising new challenges on privacy and data security, which are important issues that often come to discourage the adoption or the development of certain projects.

To address those challenges we propose *Genesy*, an innovative blockchain platform which acts as an intermediary between the owner of the genomic data and its potential users (i.e., university research centers, private laboratories, hospitals, geneticists, etc.). Our proposal envisions the use of blockchain, cloud computing and artificial intelligence as means to structure a new ecosystem that ensures the user's exclusive property and access to their genomic data, but also the possibility of participating in the benefits of the genomic research.

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.

[1] Accessible online at https://www.ibm.com/blockchain/hyperledger

2.1 Design Principles

The design principles we propose for the platform to act as an intermediary between owners of genetic data and its users can be exemplified as follows.

Design Principle 1 - Whole Genome Sequencing information sharing for custom recommendations. The value of genomic data lies in the possibility of identifying associations between genetic variants and diseases. Risk factors and strengths can be identified in advance and then, prevention, diagnosis and treatment of diseases can be targeted to achieve the best effect on each individual with a particular genetic makeup. The advantages are two-fold: (1) as for the national economy, it is about the possibility of significantly improving the productivity and efficiency of health-care and social well-being. Indeed, if genomic data will be shared with researchers, it can help to identify the causes of multiple diseases and contribute to the development of new drugs; (2) on an individual level, it means improving the quality of life and increasing the possibility that people have a direct impact on their health.

Anonymous data will be shared and processed with advanced tools for computer science and scientific analysis, also taking advantage of Big Data technologies. This way, access to data is complete and interpretations of genomic data can be continuously updated. Conditions that decree the imminent adoption of this approach.

Design Principle 2 - ecosystem access mechanisms. Users will be able to receive and share results through controlled access mechanisms and download information from Genesy locally. Users will analyze and interpret their genomic data in complete autonomy through an ad-hoc application and a service will regularly notify them of the possibility to acquire new reports applied to their DNA, which will be produced on a par with the development of new genetic panels and the progress of medical research. Access to data from users and professionals/academic or pharmaceutical facilities will take place with identity management guaranteed by private-public key pair systems and cryptographic functions. In summary, Genesy will act as an intermediary between the owner of the data and the potential users, but also as an ecosystem manager, ensuring the ownership and exclusive access of users genomic data.

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]Accessible online at: https://www.stellar.org/

compensation linked to the analysis performed on the platform, there is the possibility of attracting the attention of pharmaceutical and research companies interested to certain DNA. They will never see the name of the user: Genesy will automatically notify the user and let her/him to accept whether or not to contact these organizations for further analysis and revenues.

Design Principle 4 - transparency and privacy. The system will be designed to guarantee the maximum transparency regarding what will be shared and what will be protected by privacy, leaving the user the possibility to manage it personally (partially or totally). Users will never be forced to share their DNA without their consent. The DNA will be stored on our servers and we will be the only ones able to connect the name of the users to their DNA. However, users may decide to share it anonymously with external companies, in which case those companies will be able to query them for in-depth analyses, but ownership will remain of the users.

Furthermore, the system will be implemented and bound to follow the provisions of the "Global Alliance for Genomics and Health" for responsible sharing of genomic and health data, so to minimize damages from data sharing and maximize benefits for those contributing with their own genome, but also for societies and health systems as a whole.

In summary, our proposal aims to build a new genomic ecosystem based on the belief that all people should (i) have possession and free access to their genomic and health data; (ii) be able to control who accesses their data; (iii) be sure that the genome is safely stored; (iv) be able to improve their health in the future using their data; (v) have the opportunity to anonymously donate their data for the public good; and (vi) be able to benefit economically from the use of their genome by third parties.

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.

- **Database** - in which to store the results of the analysis and management of sequenced DNA as well as off-genomic data such as user information.
- **Smart contract** - Transactions among "ecosystem users" will be regulated by smart contract. Transactions and meta-data will be encrypted and stored on the blockchain, ensuring immutability and security. The IBM's Hyperledger Fabric[3] is the ideal tool to achieve the levels of reliability and security we require. The blockchain solution on IBM Hyperledger technology provides very high standards of security and reliability, allowing us also to integrate innovative smart contracts and our digital currency.

3 Conclusion

The new frontier of medicine is personalized medicine, where doctors will be able to recommend the most effective medicines based on our DNA. More and more advanced technologies will make our DNA a priceless mine of information. We can now analyze our DNA in an affordable cost, and soon we will be able to discover more and more about our past, present, and future. The more DNA will be analyzed, the faster the development will be. We will not only help ourselves, but we will do our part by contributing to the development of science.

The solution we outlined in the previous pages combines state of the art blockchain technologies in a new way to facilitate the acquisition of genomic data and to incentive owners of genomic data to share them to take advantage both in terms of rewards (through ad-hoc coins) and health.

A proof-of-concept of our proposal is currently under way of prototyping. A dedicated blockchain platform on IBM Fabric Hyperledger has already been implemented, and smart contracts have been developed for sequencing data acquisition and management procedures. Cloud off-chain files can be stored on IBM systems and the pipeline (in cloud) that uses raw DNA data for reporting to users has been completed. In addition, the first reports were created, highlighting the first somatic traits and predisposition to certain diseases and neurological conditions.

We are developing and integrating in the platform a series of software applications for the analysis and management of the raw data coming from the DNA sequencing activities. In terms of procedures and algorithms for the identification of genetic variants and definition of gene panels, we have implemented a standard "Genome Analysis Toolkit" environment both locally and in the cloud, which is adopted by the majority of genetic service providers. We work in partnership with the National Center for Biotechnology Information genomic data platform of the US National Institutes of Health and following the guidelines of the American College of Medical Genetics and Genomics.

We plan to complete and quickly test the platform for the management and graphical and tabular display of the users genomic information, as well how to create a network that includes "interested" geneticists and researchers who can contribute with evaluations or suggestions to the construction of a medical/scientific framework for the project. Finally, we also plan to create an internal encyclopedia based on the most important and reliable genomic databases, which will allow us to automate the preparation of our reports through the use of new DNA sequencing algorithms and their impact on research.

[3] Accessible online at: https://www.hyperledger.org/projects/fabric