Massachusetts Institute of Technology
Media Lab's Digital Currency Initiative
Sloan School of Management

Traceability in Genetic Data Sharing

Authors: Stephanie MacConnell (MIT), Nick Hong (MIT), Ajeet Singh (Harvard), Mayank Aranke (Harvard), Connery Noble (MIT)

Keywords: Healthcare applications, Traceability, Genetic Data Sharing

Blockchain Lab Program
2018-2019 Working Groups Cycle
Taught by N. Narula, G. Gensler, S. Johnson, and M. Casey
Member Company: Boston Consulting Group (BCG)
Project Group: Healthcare Applications

Traceability in Genetic Data Sharing

DCI Healthcare Working Group

Stephanie MacConnell, Nick Hong, Ajeet Singh, Mayank Aranke, Connery Noble

Table of Contents

Executive Summary

Over the past decade, significant breakthroughs in gene sequencing has allowed gene analysis and therefore data to become readily available. However, with an influx of genetic data, another problem arose: the problem of data stewardship and governance. As of today, an individual donor who has his DNA analyzed through 23andMe or Ancestry.com has no way of knowing where the data goes and for what purposes the data is used for. We believe that a blockchain solution, leveraging non-fungible tokens, will enable an individual donor to become a better steward of his own genetic data by enabling him to trace where the genetic data goes, and will strengthen the privacy of the individual's identity. By doing so, we hope to peel off any stigma around the donation of genetic data, regain the trust of individual donors, and thereby empower a community of citizens to adopt the practice of donating genetic data for the greater good of society.

I. The Problem

Currently, there are two dominant reasons a person may have provided their genetic data to a 3rd party. The first reason is for a medical diagnosis they may have needed. The other, is from personal interest, where a person has volunteered their genetic data in order to learn about one's background or health markers.

In the case of personal interest, most individuals will readily recognize the names 23andMe and Ancestry.com, both of which are direct-to-consumer genetic genealogy testing companies that provide a range of information from geographic ancestral lines to notable health indications. In the year 2017, the number of consumers purchasing geneology tests more than doubled, leading to over 12 million genomes sequenced by the end of the year through these services.[1] Subsequently, in 2018, consumers purchased just as many DNA tests as were purchased in all of 2012-2017 combined. By the end of November, 2018, Ancestry.com had 14 million genomes, with 23andMe closely behind at 9 million genomes.[2]

When consumers or patients share a sample of their genetic data to be sequenced, the collector or lab requires the patient to sign an agreement stating what can be done with the data. 23andme, for example, states that "Giving consent... means that you agree to let 23andMe share your de-identified individual-level data with approved researchers outside of 23andMe… [which may] range from academic institutions and non-profit organizations to pharmaceutical and diagnostic companies."[3] Last summer, they announced a partnership with GlaxoSmithKline, a pharmaceutical giant, to aid in the development of new medicines using 23andMe's repository of genetic data in exchange for a $300 million investment.[4] This was not their first major

[1] https://www.technologyreview.com/s/610233/2017-was-the-year-consumer-dna-testing-blew-up/
[2]https://www.cnbc.com/2019/02/12/privacy-concerns-rise-as-26-million-share-dna-with-ancestry-firms.html
[3]https://www.23andme.com/about/individual-data-consent/
[4]https://mediacenter.23andme.com/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/

collaboration with pharma- 23andMe had already partnered with companies like Pfizer, Genentec, and others by 2015,[5] and launched in 2009 what is now the world's largest genetic study of Parkinson's disease.[6] Furthermore, last year they received FDA authorization to test for genetic risk factors for 10 diseases, such as specific BRCA gene mutations associated with breast cancer.[7]

23andMe is quick to refer the public to their updated list of scientific publications, white papers, and conference presentations on their site,[8] and consumers may even be happy to know that their saliva may be supporting a greater cause. However, this fails to capture the full extent of how their data and data from similar companies are shared and used.

One customer was denied life insurance coverage because of a positive BRCA gene mutation. However uncomfortable it was for her to learn that the Genetic Information Nondiscrimination Act (GINA) did not apply to life insurance, disability insurance, or long term care, she was more surprised that such genetic tests would be used against her in the first place.[9]

Last year, the Golden State Killer was famously identified 40 years later with the help of a small genealogy site.[10] Though this was a resounding success for law enforcement, such media attention helped to hide other cases where law enforcement used genetic information to frame innocent individuals for murders they did not commit.[11] With the proliferation of DNA testing services, such companies long denied collaboration with law enforcement, but FamilyTreeDNA in Houston was recently ousted for sharing its database of over 2 million

[5] https://blog.23andme.com/23andme-research/23andmes-recent-research-collaborations/
[6] https://www.23andme.com/pd/
[7] https://www.fda.gov/news-events/press-announcements/fda-authorizes-special-controls-direct-consumer-test-reports-three-mutations-brca-breast-cancer
[8] https://research.23andme.com/publications/
[9] https://www.fastcompany.com/3055710/if-you-want-life-insurance-think-twice-before-getting-genetic-testing
[10] https://arstechnica.com/tech-policy/2018/04/genealogy-websites-identify-rape-suspect-who-eluded-police-for-40-years/
[11] https://www.themarshallproject.org/2018/04/19/framed-for-murder-by-his-own-dna

genetic records with the FBI during 2018,[12] and it adds to growing suspicions against other genealogy services of doing the same. Given America's long history of discriminatory practices by law enforcement, adding genetic capabilities enhances their capacity for unethical and unlawful surveillance and intimidation. Already, Canadian law enforcement has been using such Ancestry services to assist in identifying the nationalities of migrants and facilitate their deportation.[13]

In its more scientific contributions, 23andMe insists that its research "does not constitute research on human subjects," because it was performed on "anonymized data with no contact between investigators and participants" as demonstrated by their Institutional Review Board (IRB) application.[14] IRBs are created for the purpose of protecting the rights and welfare of human research subjects,[15] so this attempt to evade higher standards of privacy and security is unsurprising, and when genetic data can no longer be considered de-identifiable,[16] it becomes that much more important that genetic data be held to the highest standards of privacy.

In 23andMe's complete privacy statement, they offer the "Right to be Forgotten," giving all customers the right to "delete their accounts at any time",[17] but they simultaneously cite compliance with the Clinical Laboratory Improvement Amendments of 1988,[18] which, as they state in an email, means that their labs "will retain your genetic information and a randomized identifier on their secure servers for a limited period of time, 10 years pursuant to CLIA

[12] https://www.nytimes.com/2019/02/04/business/family-tree-dna-fbi.html
[13]https://www.reuters.com/article/us-canada-immigration/canada-using-dna-ancestry-websites-to-investigate-migrants-idUSKBN1KH2KF
[14] https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000993
[15] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631034/pdf/chest_148_5_1148.pdf
[16] https://science.sciencemag.org/content/339/6117/321
[17] https://www.23andme.com/about/privacy/
[18] Clinical Laboratory Improvement Amendments (42 USC 263a) https://www.govinfo.gov/content/pkg/USCODE-2011-title42/pdf/USCODE-2011-title42-chap6A-subchapII-partF-subpart2-sec263a.pdf

regulations."[19] Before the GSK deal, a researcher at the University of Queensland[20] estimated that 23andMe made near $130 million USD from selling access to a million genotypes.[21] Today, they continue to build a handsome revenue stream, sharing data with profitable pharmaceutical companies, using rich genetic data that consumers paid them to provide, while consumers are given limited visibility into how else their genetic data will be used.

THIS IS FUNDAMENTALLY A CHALLENGE OF TRUST

In the case of medical diagnoses, select premier hospitals have built into their workflow a standardized process for consenting patients for the collection of genetic information. For example, at Memorial Sloan Kettering Hospital, a leading cancer center in New York, every patient is sequenced as part of their care plan (Juan Perrin, interview, April 18, 2019). Alongside them are similarly renowned institutions such as Brigham and Women's Hospital in Boston and the MD Anderson Cancer Center in Houston.

An interview with a leading Memorial Sloan Kettering researcher revealed that they exclusively share data with faculty and analysts inside of the labs at their own institution. We found a similar story echoed across institutions, that patient data is generally de-identified and shared for specific purposes to third parties who request either specific data for ongoing research or, earlier in the process, for general understanding of population numbers around an interest area.

This data sharing can be a great revenue generator for organizations. Dermot Shorten, Quest Diagnostic's Vice President of Strategy and Ventures, said his company earns in the low single digit millions per year from such sales.[22] However, through our research, we

[19] https://www.bloomberg.com/news/articles/2018-06-15/deleting-your-online-dna-data-is-brutally-difficult
[20]https://www.reuters.com/article/us-health-dna/cashing-in-on-dna-race-on-to-unlock-value-in-genetic-data-idUSKBN1KO0XC
[21]https://www.reuters.com/article/us-health-dna/cashing-in-on-dna-race-on-to-unlock-value-in-genetic-data-idUSKBN1KO0XC
[22] http://clinchem.aaccjnls.org/content/clinchem/early/2016/12/29/clinchem.2016.261479.full.pdf

discovered there is no centralized way to buy and sell data in the current system. Seven Bridges Genomics,[23] was trying to address this challenge, but described it as "complicated," where the vast majority of transactions were left to a manual process. This process is network-driven, with for example a pharmaceutical company's director reaching out to an executive at a lab that processes genetic data. They will email back and forth or schedule calls to kick off the process, and the pharmaceutical company will define the information they are looking for. The lab with then either query their own data to provide an answer or depending on the engagement will transmit the data to the buyer. Another initiative to address this pain point is cBioPortal,[24] which provides large-scale cancer datasets, analysis and visualization. It was originally developed at Memorial Sloan Kettering and is available via an open source license. There is little in the way of a standard transmission format, historically it has been an HL7 version 2 interface specification,[25] but those interviewed agree that this format hasn't taken things like genetic testing strongly into consideration. Once the data is gathered, pain points were uncovered through multiple interviews around storage of the data. Traditional databases are being used, with industry leader Quest Diagnostics mentioning almost all is held in isilon storage, a type of archival storage, with hundreds of petabytes being accumulated over time. Quest Diagnostics is one of the labs for direct-to-consumer testing company Ancestry.com.

All interviews confirmed that outside of medical diagnosis delivery, there is no feedback loop to the patient about what happens with their data, where it is sold, or what it is being used for. An interviewer stated, "It would be very valuable to have this information. The more metadata, the better, even if we don't see relevance now - where it went, what was done with it. Especially for sensitive stuff like DNA - even after being cleaned, this data is still identifiable." An

[23] https://www.sevenbridges.com/
[24] http://www.cbioportal.org/
[25] HL7's Version 2.x (V2) messaging standard is the workhorse of electronic data exchange in the clinical domain and arguably the most widely implemented standard for healthcare in the world. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=185

interesting point around consent revealed that because of tight language and regulation, even if outliers in a patient's genetics make themselves known, a medical lab can generally not tell them if diseases are diagnosed or identified if that issue is not what they specifically consented to.

This profound lack of transparency for patients has contributed to an even greater challenge pressing the medical community. A study published in Genome Biology last year showed how our cutting edge genome-wide association studies (GWAS) failed to appropriately represent diverse ancestries. GWASs look for small variations in the genome to isolate and identify specific genes that may contribute to a patient's risk of developing a given disease.[26] From all GWAS publications- the 4,600 studies included in this assessment- 78% of participants represented individuals of European descent, with Asians comprising 11%, and Africans and Latinos combined forming <4% of study participants.[27] As a result, much of the research driving the entire discipline of precision medicine- "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person"- is failing to adequately represent our diverse population.[28]

One can easily imagine a possible genetic mutation prevalent in the Black American population that changes the way certain molecules are metabolized in the body. A pharmaceutical company, relying on GWAS insights, may invest hundreds of millions of dollars to develop a new drug, but not realize until clinical trials go live that a particular mutation leads to a rapid buildup in the body, leading to serious harm or even death of Black individuals. Unless the data used is equitable in how it represents the population it aims to help, whole communities will be left out of the innovation cycle, and the research findings may lead to inappropriate diagnoses, treatments, and even patient harm.

[26] https://ghr.nlm.nih.gov/primer/genomicresearch/gwastudies
[27] https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1396-2
[28] https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

Minority populations have long been distrustful of healthcare research and treatment services. In the early 20th century, the Tuskegee Institute and US Public Health Service notoriously engaged in a treatment study of Syphilis, but the African American population was intentionally not given any penicillin treatment, and were instead observed over 40 years to study the effects of the treatable disease.[29] Even at the turn of the century, peer reviewed studies of African American populations were found to be less trusting than white Americans regardless of their social class.[30] As it stands, they are already underrepresented in research studies for conditions like prostate cancer, for which African American men are already more likely to develop and die from than Caucasian men.[31]

The historical lack of accountability coupled with the modern lack of transparency in genetic data collection makes it nearly impossible to build trust and incentivize participation in medical research. As long as these key issues in trust and transparency are not addressed, minority populations will not believe there will be accountability, and continue to resist opportunities to drive medical research and advances in treatments for their populations.

[29] https://www.cdc.gov/tuskegee/timeline.htm
[30] https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/214437
[31] https://www.auajournals.org/article/S0022-5347(17)30779-6/fulltext

II. The Solution

A. Current Developments

1. Non-Genomic Solutions in Permissioning and Data Ownership:

The challenges with healthcare data ownership and exchange are hardly unique to the genomics space. The issues elucidated[32] above are seen in several other areas of the healthcare ecosystem, and there have been several efforts to try and make the transaction of healthcare data more efficient. The use cases below illustrate a few examples of companies working to make data exchange more efficient and democratic in both, the genomics space and other areas of healthcare. Reviewing the current state of such solutions and the problems they are trying to address is helpful in understanding where significant innovation is happening, and consequently, where opportunities for paradigm shifts exist.

SimplyVital

To understand SimplyVital's solution offerings more fully, a brief discussion of value based payments and care coordination is useful:

The healthcare industry is seeing a shift in payment patterns from fee for service payments to value-based care programs, which shifts focus from quantity of care (number of procedures performed) to patient outcomes from the care they've received. This shift requires a high degree of coordination between several providers that a patient interacts with during a said episode of care.

The successful recovery of a 55-year-old, overweight male who has had to undergo a complete knee replacement, for instance, involves several care providers- the orthopedic

[32] To make clear or plain, especially by explanation; clarify.

surgeon, the pharmacist, the physical therapist, even the patient's primary care provider. Communication between these providers, while crucial for successful patient recovery (in turn, for provider reimbursement under the value-based care model), is tremendously inefficient.

Providers use disparate electronic medical record systems, fax machines, excel spreadsheets and other disjointed modes of communication to convey different episodes of care that the sample patient above might have gone through, and information that is lost in these inefficiencies leads to large cost burdens. One report by Accenture estimates that hospitals waste $12B annually due to poor communications. Processes or tools that increase communicative efficiencies are therefore significant value adds in this value-based care framework that financially incentivizes sharing of medical data amongst providers.

SimplyVital's *Sauna* solution aims to solve the problem of inefficient data exchange between disparate providers by storing episodes of care as data points on the blockchain. Care coordinators and providers that are associated with a said patient's care are notified every time the patient receives care from a provider within their care management team.

Subsequently, transaction receipt of both, the episode of care taking place, and the care coordinators and providers getting notified, are hashed and stored on the blockchain. Providers can then use *Sauna* and SimplyVital's underlying Ethereum based protocols (HealthNexus) as an audit trail for when episodes of care happened, and subsequently use this for reimbursements in the value-based care model.

Sauna illustrates a key example of breaking down data silos and decentralizing/democratizing an archaic model of data ownership. SimplyVital has another solution, *Agora*, that builds on the network of stakeholders that the company plans to build using their *Sauna* solution and is focused on improving data accessibility. This product is still in its early stages, but from an interview with the company, their focus is to simply make data

accessible for buyers and sellers- SimplyVital is agnostic to the type of data that will eventually be transacted (although they concede that genomics and oncology are intriguing)[33][34]

Agora operates on the premise that it will act as a third-party intermediary/marketplace (think Amazon) and simply connect parties that have stores of data- presumably providers- to parties that desire data- i.e. pharmaceutical companies for research or insurance providers for usage trends. All transactions on *Agora* are designed to take place using SimplyVital's own token- *HealthCash*- and the proposed marketplace is planning on incorporating fiat to crypto exchange capabilities[32].

Though SimplyVital's solutions do not shift the current data ownership paradigm- the providers are still owners of health data- its solutions do provoke thought on blockchain based permissioning[35] and its role in creating an efficient exchange of data. *Sauna* illustrates how provider permissioning between trusted stakeholders can drive increased revenue in the value based care model and *Agora* shows the same concept can enable a marketplace for medical data.

MedRec

MedRec was a whitepaper prototype published by the MIT Media Lab and BIDMC in 2016. Where SimplyVital's focus is on provider-provider and provider-industry data sharing, MedRec brings patient agency over their own data to the center of its proposed solution. To understand MedRec's value proposition, a quick review of the dominant issues with electronic medical records (EMRs) is helpful. It is widely agreed upon that EMRs in the current state:

[33] https://crushcrypto.com/wp-content/uploads/2018/03/HLTH-Whitepaper.pdf
[34] Connery + Stephanie Interview With David Akers (SimplyVital)
[35] Permissioning is defined here as an efficient and safe way for one party to grant another party access to information that the former party is holding. This can be for purposes of information validation (occurrence of an episode of care) or for transaction of value (HealthCash traded for medical data).

1. *Fragmented and slow access*- Current EMR design promotes fragmented storage across multiple healthcare settings as patients change providers or move geographically. This results in loss of past medical data as the patient is no longer the steward of their own data[36]. HIPAA privacy rules also promote a slower exchange of data, allowing providers up to 60 days to respond to requests for access (and edits to) past medical records.
2. *System interoperability*- the current disparate EMR system lacks any incentives for EMR companies to design frameworks that facilitate data transfer from one system to another. In fact, it can be argued that preventing such transfer actually result in significant benefits for EMR companies[37].
3. *Lack of availability of research data*- it is notoriously difficult and time intensive for researchers to mine EMRs for studies. Further, patients, care providers and regulatory agencies have noted an increasing interest in wanting to contribute to research, but currently lack efficient ways to do so[38].

MedRec's suggested prototype proposes a solution to these problems by employing a private blockchain where the block content represents data ownership and viewership permissions determined by members of a peer-to-peer network. They also employ Ethereum's

[36] 1. "Who Owns Medical Records: 50 State Comparison." Health Information and the Law. George Washington University Hirsh Health Law and Policy Program. Aug. 20, 2015. [Online] Available: http://www.healthinfolaw.org/comparative-analysis/who-owns-medical-records-50-state-comparison

[37] https://www.healthit.gov/sites/default/files/reports/info_blocking_040915.pdf

[38] Kish, Leonard J., and Eric J. Topol. "Unpatients [mdash] why patients should own their medical data." Nature biotechnology 33, no. 9 (2015): 921-924

smart contracts[39] to create representation of medical records on individual nodes, and these contracts in turn contain metadata on record ownership, permissions and data integrity[40].

Several other cryptographic properties build out the proposed prototype and result in a theoretical system that offers patients a decentralized, immutable record of their health data across providers and treatment sites. This system also allows providers and patients to securely offer research entities access to their health data (anonymized and aggregated).

While this model is different from SimplyVital's vision of a healthcare data marketplace, it illustrates a similar principle of how Blockchain can enable secure patient and provider permissioning to enable more robust clinical research. More importantly, the MedRec prototype offers a distinct change in the current data ownership model; patients are now stewards of their own data, and not hospitals and providers.

2. Incentives focus on monetary or personal gain

It is yet unclear as to what will truly motivate customers to donate their genetic data. But one thing is clear: the lack of privacy and traceability on genetic data is having people start to question or even regret their decision to give away genetic data.[41] In an effort to resolve this problem where donating genetic data is becoming a stigma, a number of companies are trying to provide the right solution to incentivize patients to give genetic data. To name a few, Foundation Medicine, LunaDNA, and Nebula Genomics are active incumbents in the space. But first, it is imperative to understand how these companies operate.

Foundation Medicine, LunaDNA, and Nebula Genomics all have similar revenue models. They each have a database of customers' health data. Commercial organizations such as

[39] Smart contracts are self-executing contracts with the terms of the agreement between buyer and seller being directly written into lines of code. The code and the agreements contained therein exist across a distributed, decentralized blockchain network. https://www.investopedia.com/terms/s/smart-contracts.asp

[40] https://www.healthit.gov/sites/default/files/5-56-onc_blockchainchallenge_mitwhitepaper.pdf

[41] https://thewestsidestory.net/the-23andme-deal-with-glaxosmithkline-raises-privacy-and-ethics-questions/

pharmaceutical companies, biotech, laboratories, and other institutional organizations pull queries from these databases for medical research, which then can be used for drug development. These queries are revenue-generators for Foundation Medicine, LunaDNA, and Nebula Genomics. And the following three companies want to attract more customers who donate genetic or health data because more data creates a positive reinforcement for commercial organizations to come back to pull queries. Data is each of the company's unique competitive advantages. What is different amongst all of the three companies however, is their approach in attracting customers to donate genetic data.

Foundation Medicine connects physicians and patients to their cancer treatment approaches under the consent that the patients' data is given to Foundation Medicine for analysis. The patients in return are given recommendations on potential treatment or appropriate drugs depending on their disease. In this business model, patients have the incentive to give away data because their donation can directly help them find proper diagnosis and treatment.

LunaDNA is the first community owned health data platform where patients can anonymously share health data on one platform to advance medicine. Every member of LunaDNA who shares their genetic information receives shares of LunaDNA, giving them ownership. Commercial organizations such as pharma or biotech will pull queries from LunaDNA's central database storing health data, which in return generates revenue for LunaDNA. All proceeds after covering for expenses, are returned to shareholders who are members that have donated their data. LunaDNA envisions customers becoming true stewards of their own data, and giving back to society as part of a community, which is also known as "citizen science".

Nebula Genomics is exploring ways to incentivize customers by focusing on privacy, transparency, and micropayment. When customers get their whole genome sequenced through Nebula kit, customers will receive raw genetic data, and have access to a Nebula blockchain

that stores the genetic data, and receive an ancestry report. Also, they are looking to shift sequencing costs to researchers and enable people to get compensated for data sharing.

Foundation Medicine, LunaDNA, and Nebula Genomics touch upon different elements of incentive to attract customers who are willing to donate genetic and health data. It remains to be seen what will truly motivate customers and what incentive is executable, but we believe there is a different way to incentivize customers while improving their data's traceability and privacy in a more effective way.

B. Our Proposal

1. Citizen Science

We propose adjusting the workflow using blockchain technology. The two pain points most readily addressed by distributed ledger technology are the lack of incentive for genetic data donation and the traceability of the data following donation. Our proposal calls for using zero knowledge proofs and non-fungible tokens (NFTs) in tandem, which has been done less than a handful of times before, to unlock the potential for a new, secure feedback loop to the donor that can serve as an incentive to contribute to a citizen science community.

NFTs and Citizen Science

The first blockchain "game," CryptoKitties, introduced the idea that consumers might pay for a digital item that has meaning to them, that the consumer could track. This model was put to use in the context of citizen science for the first time during an experiment in July, 2018. Axiom Zen, the company behind CryptoKitties, partnered with NGOs Ocean Elders and ACTAI Global to create "Honu", a sea turtle-inspired CryptoKitty. The groups put the digital sea turtle up

for auction to raise money for sea turtle conservation efforts and sold for $25,000.[42] This is a notable trend because it marks the first time the larger market has embraced what has been described as essentially tokenizing “DNA” onto the blockchain to identify specific assets (cats in this case.) A non-fungible token is a special type of cryptographic token which represents something unique; non-fungible tokens are thus not interchangeable. This is in contrast to fungible tokens like bitcoin, and many network or utility tokens that are fungible in nature.

Another example use case of NFTs is David Noble’s GunClear product, which he stated in an interview was inspired by a desire to take the idea of CryptoKitties and build thick layers of privacy around it. This idea is the first we are aware of to use NFTs and zero knowledge proofs together. In this use case, the sale of a firearm from one party to another is made more secure by having a clearly traceable history of gun ownership and tokens can be transferred to signify a sale, but with privacy for the buyer and seller. The only thing public in the interaction is the gun’s serial number.

Proof of Impact (PoI) is yet another unique use case that is aiming to change the current state of social capital markets by using NFTs. PoI has several financial investment models tailored to individual use-cases, but all are built on the premise that investors (buyers) can purchase impact events- real world actions that range from childhood vaccinations to carbon credits. When a sellers performs an impact event, they deliver data to prove the completion of the impact event which is then verified on the public ledger, then an NFT corresponding to this impact event is created and awarded to the seller. These unique impact event tokens can then be sold to buyers. . The buyer therefore has full visibility into the details of the impact event they contributed to and the real world event it helped affect.[43]

[42] https://news.mongabay.com/2018/07/how-blockchain-gaming-could-benefit-wildlife-conservation/
[43] *Proof of Impact Whitepaper: Unlocking the Intrinsic Value of Impact through Global Impact Capital Markets*, March 2019
https://drive.google.com/file/d/1gDqxorF_le4WU1TNt8RYWGqkp23EWqQJ/view?usp=sharing

Exploring the latest use cases of these technologies outside of healthcare has helped to more clearly define the potential for novel workflows within the industry. For genetic data, introducing the ability for a donor to know exactly where their data went - to support research, contribute to which clinical trial, study which disease - and further, the ability to one day permission which of these the donor does or does not allow, can be an incentive to donate in the first place.

Transparency/Traceability

One of the key aspects that hasn't been dealt with by the current incumbents is the traceability aspect of customer genetic and health data. In current practice, once genetic or health data is given to a B2C company or a healthcare provider, a customer is either promised micropayment or protection of identity; however, no customer really knows how the genetic data is used, and for what purposes pharma or biotech companies pull queries that pertain to his data.

We believe that by having the customer's genetic data transacted through the blockchain, we can provide traceability to the customer-- a consistent feedback loop where the customer is alerted or notified of the specific ways his genetic data has been used. By doing so, customers not only feels secure, but also can engage more in the process of advancing a medical cause. A simple way to understand this situation is to think of sharing an online photo at a major social network. Imagine Person A (Alice) posts a group photo on her profile page. Person B (Bob) who also happens to be Alice's friend, decides to share the image that Alice posted. Alice may be notified. Then, Charlie who is Bob's relative, sees the post on Bob's wall, and decides to share the photo again. Unfortunately, Charlie is not an acquaintance of Alice. Therefore, Alice is completely unaware of Charlie sharing her photo, and has no way of finding out how her photo has been exposed to a completely new network.

We believe that the first step to enabling customers to become stewards of their data is by improving traceability and privacy. Fortunately, blockchain provides the technical backbone to bring improvements to fruition.

2. Technical Details

In this proposal, we outline a high level system concept. We highlight the key technical requirements to satisfy the system, while some of the more detailed technical specifications are left to be determined by specific business, implementation, or roll out requirements.

Traceability and Privacy via NFTs and Zero Knowledge

The primary idea behind this proposal is to allocate a unique non-fungible token[44] (or NFT) as an identifier for a specific piece of genetic data. Recipients of the genetic data can then use the NFT as a reference, in order to append a log of activities in such a way that the original owner of the data can trace the activity, without needing to know any actual information about them.

The NFT is effectively the focal point for logging and tracing of how a piece of data is used and passed through the system. Being able to trace where and how data is used is a critical component to this proposal, however, only the original data owner should be capable of viewing the trail of their data. Therefore, in order to ensure the privacy of this process, Zero Knowledge Proofs[45] should be used to hide the true value of each NFT, so that it remains private between the parties.

[44] A non-fungible token (NFT) is a special type of cryptographic token which represents something unique; non-fungible tokens are thus not interchangeable. This is in contrast to cryptocurrencies like Bitcoin, and many network or utility tokens that are fungible in nature.

[45] Zero Knowledge Proofs are used when two parties want to prove that they know some information without revealing the value itself. There is often a need to perform authentication without exchanging passwords, which also means that the passwords cannot be stolen. The term “Zero-Knowledge” stems from the fact that zero information is revealed, but the verifying party is correctly convinced that the proving party knows the secret.

The exact implementation of Zero Knowledge Proofs can vary depending on specific requirements. It could be as simple as a hash[46] involving the NFT and message, which clients can recompute to verify knowledge of the NFT. Or, it could be as involved as leveraging zk-SNARK[47] to prove authenticity before a message is even accepted onto the blockchain. Eitherway, the NFT value should remain private, and should never be publicly disclosed on the blockchain.

Furthermore, any time the data is shared with additional parties, a new derivative token should be generated to ensure that each party is able to transparently report their uses of the data, without being able to themselves trace who and how the data is being used by others. Only the original token owner should be able to reconstruct the full trail of their data.

Both the Zero Knowledge Proofs and derivative tokens are important for two main reasons. First, we want to ensure that the messages are authentic. That is, they originate from a recipient who does indeed have the actual data (or, rather, the NFT). Second, 3rd party observers should not be able to reconstruct or correlate the flow of any piece of data. Although the NFTs do not, in themselves, have any information that can identify an individual, data in aggregate can start to infer additional information associated with the data.[48] Therefore, we want to ensure that only the original data owner is provided with full traceability of where and how their data is used.

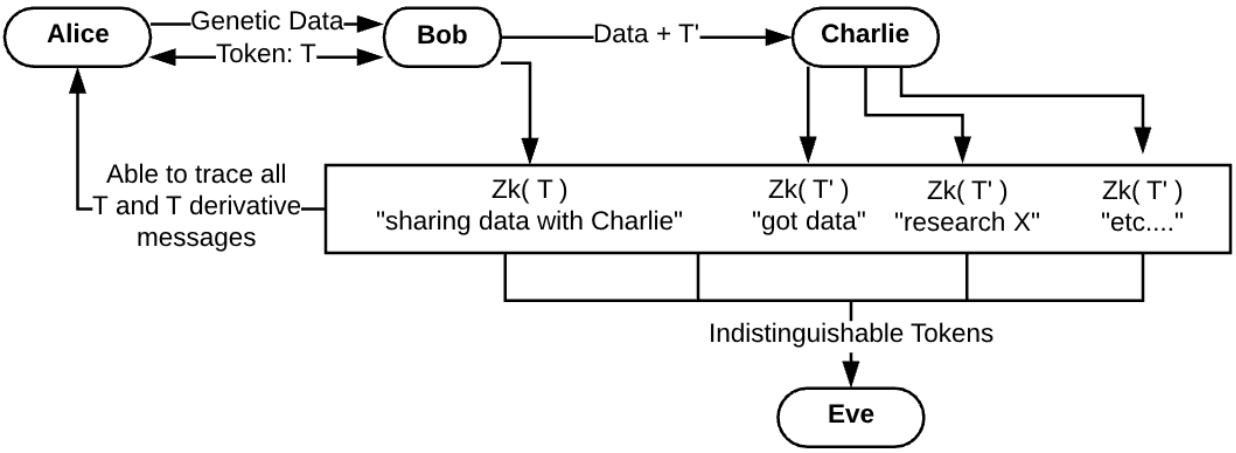
[46] A one-way function designed to convert any data into an output of a fixed size.

[47] zk-SNARK stands for "Zero-Knowledge Succinct Non-Interactive Argument of Knowledge," and refers to a proof construction where one can prove possession of certain information, e.g. a secret key, without revealing that information, and without any interaction between the prover and verifier. https://z.cash/technology/zksnarks

[48] L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Alice and Bob[49]

Alice gives her genetic information to Bob, which is then allocated token T. Bob can log events for Alice to trace via Zk(T). When Bob shares data with Charlie, he generates another token T' for Charlie to use. Charlie can log events for Alice to trace via Zk(T'). However, Charlie cannot trace any of Bob's Zk(T) messages (or any other tokens bob may have given to other partners). Meanwhile, Eve is watching the whole chain but cannot distinguish or correlate any of Bob's or Charlies messages.



Why Blockchain

We believe the use of a blockchain is crucial to the success of this proposal. Although the interactions involved between parties and the actual data being stored is simple, the need for a delicate balance of transparency and privacy can only be satisfied with a publicly visible and verifiable ledger The concept proposed in this paper does not require some of the more traditional strengths of a blockchain (like bizantine consensus, the double spend problem, decentralized authority/trust), however, it leverages open (unobstructed) access between

[49] Alice and Bob are the world's most famous cryptographic couple. http://cryptocouple.com/

disparate parties and emphasises a need for accountability. A traditional database cannot offer the same level of transparency, accessibility and auditability needed to make this a success.

That is not to say, however, that the system must be run on a public blockchain. The specific operation could very easily be implemented to run on an existing chain (ex: Ethereum), or as a small private blockchain.

III. Challenges/Risks

It is known the healthcare field can have many large boulders to move. The genetic data workflow has historically been very segmented and buy-in is needed from multiple large stakeholders. The value proposition must be clear and compelling. For best results, labs, research institutions, pharmaceutical companies and other stakeholders would need to adopt and maintain this blockchain. The obvious questions for further review include who will make the money in this value chain and where the incentive for the stakeholders lies.

IV. Additional Considerations

The model we present in this paper is specifically built to provide a secure platform for recording transactions of consent for sharing genetic data. As such, we are not addressing the best practices for exchanging the genetic data itself. There has been some literature produced around the appropriate infrastructure and policy for enabling the sharing of genetic data.[50] However, the industry would benefit greatly through continued discussion and collaboration, ideally with the participation of stakeholders including hospitals, researchers, and laboratories, as appropriate to the available technical capacities for the participants and their respective institutional and federal privacy requirements.

[50] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5862236/

Another consideration we chose not to detail is the importance of effective user interface design. For users, regardless if they are patients in a hospital or consumers at home, the determination of whether they actually feel empowered to control and direct the movement of their genetic data will depend less so on the technical architecture behind the product, and more so on how they interact with the tool itself. There has been exhaustive literature over the past few decades, long before the proliferation of Electronic Health Records, highlighting the need for iterative design and improvement of user interfaces, sensitive to the "information requirements, cognitive capabilities, and limitations of end users".[51] More recent reporting has revealed how non-intuitive user interfaces can actually be disempowering, invite errors and even cause harm.[52] As blockchain technologies strive to preserve the integrity of patient consent, developers need to be vigilant in the design of their tools, and ensure that users are able to engage in informed consent, where users can feel confident that their decisions were based on the disclosure of sufficient information. Respecting existing ethical and legal standards, this means that the information for users is accurate, adequate, and relevant, and the burden rests on vendors to ensure their products meaningfully demonstrate this for users.[53]

If this model were to successfully improve the diversity of available genetic information, stakeholders would need to be wary of not magnifying other inequities in how human populations are represented for research. For example, a recent GWAS from the UK Biobank identified 30 independent genetic loci, or points on a chromosome, that have been associated with household income.[54] If users who choose to join the platform come from different, previously underrepresented genetic ancestries, it will improve the ethnic diversity of data; but if among them, only the wealthy participate, there will remain in the data potential gaps in representation. Furthermore, if the user interface enables consent mechanisms for participation

[51] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232103/
[52] https://khn.org/news/death-by-a-thousand-clicks/
[53] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2840885/
[54] https://www.biorxiv.org/content/10.1101/573691v1

in individual trials of genetic research, users may be able to selectively decline participation in trials for specific diseases or pharmaceutical companies. Both of these are examples of potential selection bias. As a result, deliberate collection and study of detailed demographic data will be necessary to ensure that patterns of user participation and consent do not inadvertently misrepresent the population intended for analysis.

Policy gap: The Genetic Information Nondiscrimination Act (GINA) of 2008 was the first and only of its kind, attempting to build a "a national and uniform basic standard... to fully protect the public from discrimination and allay their concerns about the potential for discrimination, thereby allowing individuals to take advantage of genetic testing, technologies, research, and new therapies."[55] While GINA was considered a success, promising protection from discrimination in employment and health insurance, it failed to include explicit protections for the purchase of life, disability, or long-term care insurance. Furthermore, the protections only apply to those without manifest disease, suggesting that as soon as an individual begins to develop symptoms of their genetic condition, GINA no longer applies and the individual is left vulnerable to discrimination. In one ongoing study of whole-genome sequencing, 25% of the participants who declined to participate cited "fear of insurance discrimination as the primary reason, after a consent process in which they were specifically educated about GINA."[56] This presents additional barriers to participation for any genetic data marketplace, regardless of the technologies deployed, and such gaps in policy would likely require federal legislation to correct.

V. What's Next/Path Forward

If non-fungible token and the blockchain technology truly enables a customer to trace the routes of their genetic and health data, it also potentially opens doors for micropayments in the

[55] https://www.eeoc.gov/laws/statutes/gina.cfm
[56] https://www.nejm.org/doi/10.1056/NEJMp1404776

future. This is possible because a digital ledger that the customer owns has all the transaction records that trace data movement from patient to hospital or DNA collector to laboratories to pharma and biotech companies. By having transactional evidence, customers can claim payment for contribution to oncology or rare disease research. However, uncertainties still remain. For example, we do not know who will make these payments and the magnitude of the payments. We do not know whether any parties will consent to making payments to customers for that matter. Also, it is yet unclear how customer's identity privacy will be protected when micropayments will need to be traced back to the person who deserves the payment.

Appendix – Interview List

Collectors
George Church, Founder Nebula & Veritas
Kamal Obbad, CEO and Dennis Grishin, CSO Nebula Genomics
Andrew Hessel, CEO Humane Genomics
Bob Kain, CEO LunaDNA

Labs and Industry
Scott Chapin, Head of Systems Architecture Quest Diagnostics
Juan Perrin, Director Memorial Sloan Kettering
John Halamka, CIO Beth Israel
Max Duffy, Associate Director Foundation Medicine

Technical Interviews
Kat Kuzmeskas, CEO and David Akers, Senior Blockchain Engineer Simply Vital
Sophie Meralli, zkLedger

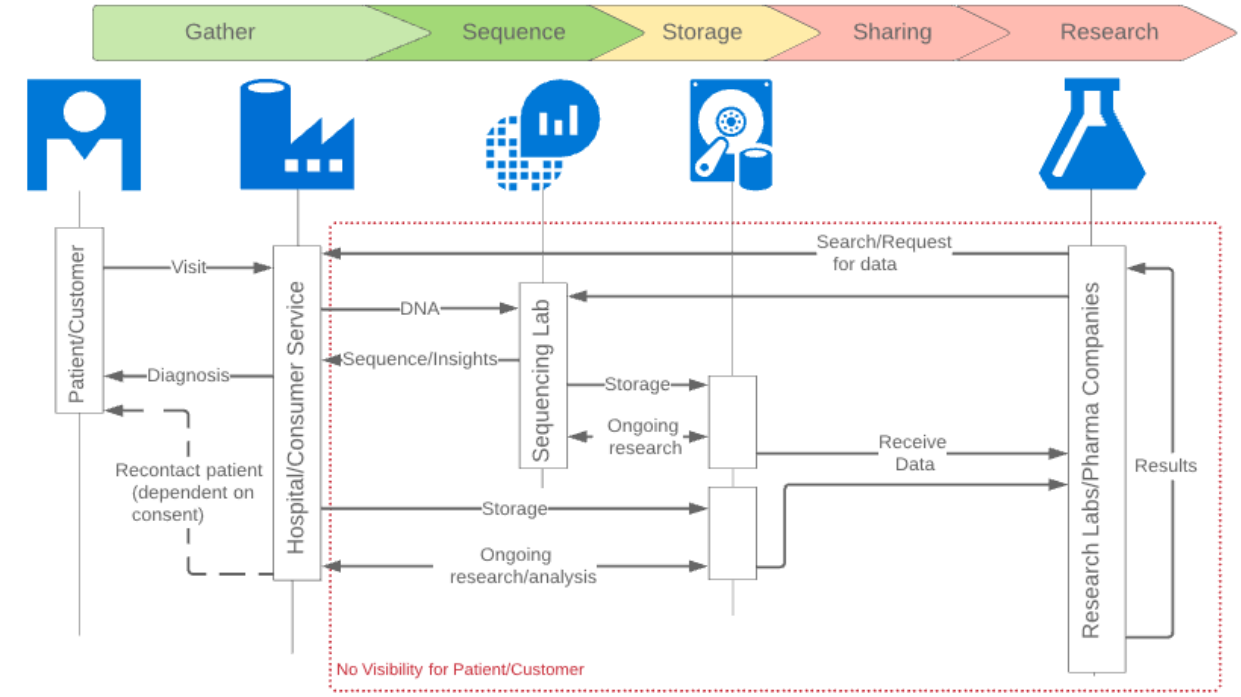
NFT and Citizen Science Interviews
Arnold Waldstein, NFT Expert
David Noble, CEO Gunclear

Appendix - Other Resources

1. A scalable, aggregated genotypic–phenotypic database for human disease variation: https://academic.oup.com/database/article/doi/10.1093/database/baz013/5316668
2. With startup, George Church bets cryptocurrency will boost genome sequencing: https://www.statnews.com/2018/02/07/george-church-cryptocurrency-genome-sequencing/
    1. Nick with the interview: https://medium.com/@nickhong89/exclusive-interview-with-george-church-godfather-of-gene-engineering-part-1-of-2-3b4ea8b59227
3. Patient Consent and the Commercialization of Lab Data http://clinchem.aaccjnls.org/content/clinchem/early/2016/12/29/clinchem.2016.261479.full.pdf
4. Here's what kind of data genetics testing companies can share https://www.popsci.com/genetics-testing-privacy
5. The value of lab data in the healthcare ecosystem https://www.beckershospitalreview.com/population-health/the-value-of-lab-data-in-the-healthcare-ecosystem.html
6. https://www.technologyreview.com/s/610233/2017-was-the-year-consumer-dna-testing-blew-up/ number exceeds 12m

Appendix - Diagrams

Current State:



Future State: