Credit: LuckyStep48/Alamy Stock Vector

Your DNA broker

Millions of people each year are giving up their genotype information for research, for health, for fun and now for profit.

Laura DeFrancesco and Ariel Klevecz

Martijn van Kalsbeek is a Dutch internet technology specialist who has dabbled in trading cryptocurrency in recent years. He is one of a small group of people exploring a completely new way of sharing their personal health data online. His millennial philosophy regarding data sharing goes against those of traditional direct-to-consumer (DTC) companies, such as 23andMe: "Whenever you donate something biological—be it blood, organs, marrow, sperm or data—for the greater good, no company involved should profit from its sales," he says. For this reason, van Kalsbeek recently took the plunge and joined a new type of data brokering service: EncrypGen is a brokerage that gives individuals control over their personal DNA data and how the data are sold to other users, researchers or companies. It is one of a nascent and growing genre of commercial ventures that aim to completely leapfrog over existing DTC genetics companies.

Some among these companies have jumped on the blockchain and cryptocurrency bandwagon, which has led to an interesting admixture of 'techies' who follow the trends in crypto-investing and researchers interested in advancing science. Daniel MacArthur, a genomics researcher at the Broad Institute in Cambridge, Massachusetts, USA, says, "It's largely a consequence that you have this neat intersection between two buzzwords—cryptocurrency on one hand and genomics on the other—and that makes for an appealing package potentially for start-up founders and investors."

However, this new approach is raising all kinds of questions for academic and industry researchers alike. Will data brokerages amass a sufficient critical mass of personal health data to make their model attractive to data buyers? Will a sufficient number of people be willing to put their faith in this completely new way of exchanging personal data? And how will the marketplace determine the value to place on personal health data?

Whom do you trust?

As technology advances make DNA sequencing ever cheaper and faster, inevitably millions more genomes, particularly human ones, will be sequenced in the coming years. "This is just the beginning. It's nothing—it's a drop in the bucket," Yaniv Erlich, bioinformaticist and chief science officer at MyHeritage recently told MIT's *Tech Review*[1]. In fact, more people got their DNA genotyped in 2018 than in all prior years combined, owing in

Box 1 | How does blockchain work?

Blockchain provides a transparent contract within a virtual environment that automatically carries out functions built within the contract or from uniformly agreed-upon interfaces. Best known in its iteration in trading currency (for example, Bitcoin), blockchain has many other applications: it provides a tool for defining ownership over something and for allowing the trade of goods to be automated without the intervention of an intermediary. In the context of managing the transfer of sensitive data, such as genome data, a blockchain would allow data owners to interact directly with data buyers.

An important feature of blockchain is that it is a transparent system that shoulders the trust necessary in valuable transactions between two parties. Hashes—computationally generated, unique strings of text used as an identifier—encrypt information and link blocks together in a way that makes the chain immutable. The hashes and encrypted information can be anything, as long as they are unique and are referred to the previous block, linking each block to the next. In theory, instead of using heavy computations to create encrypted hashes, it can be bootstrapped with DNA encoding. In fact, a popular tutorial for building blockchain 'Dapps' constructs 'DNA' in their examples. In a silly and very popular game called Cryptokitties, cats are bred as digital collectors' items; some of the rare combinations are now worth hundreds of thousands of dollars, and their appearance is based on their encrypted hashes, which are referred to as their 'DNA'.

Privacy is attained by de-identifying the data owners—who are known only by their assigned hashes—and by encrypting the data. Technologies such as homomorphic encryption and multiparty computation have been developed that allow researchers to work on encrypted data. Before these tools were developed, data were encrypted whenever they were moved and then de-encrypted while being studied. With these new tools, the data can be probed without de-encryption. However, each approach has its advantages and disadvantages: homomorphic encryption is difficult to scale, so it is slow; multiparty computation is scalable but lacks the precision of homomorphic encryption.

The space that the data buyers have available to analyze the data is completely up to the companies providing this architecture, and it does not have any contingency regarding the use of blockchain or a centralized store. Some companies (as well as database owners) fragment the data, as one way to keep the data secure. In some cases, as part of the empowerment of individual data owners that companies are purporting to enable, the data owners choose how their data are stored, including on their own equipment (though many people either will not be able to or will not want to) or on some HIPPA-compliant cloud storage. Consequently, and somewhat ironically, in many cases, megaliths such as Google or Microsoft will be where the repositories reside.

There are multiple 'chains' built by different companies (for example, Bitcoin versus Ethereum) that behave differently. Currently, Ethereum is the frontrunner as the developer space, and Bitcoin is the clunky sort of genesis that is filled with value; there are others claimed to have better standards and scaling capabilities, but they are still fighting for a substantial stake in the market.

part to aggressive advertising campaigns by the DTC genome-testing sector[1].

This genotyping could be a boon for human geneticists—if they were able to access the data. But at least for now, the genomes are scattered around the globe, and the data provenance and access are under the control of the companies and institutions where the data are generated. Whereas some public repositories (in academia and in clinical research centers) provide access to their data to credentialed researchers, the (for-profit) DTC companies have turned their genome collections into cash cows: they charge their customers for generating their genotypes and then create different revenue streams from their data stores. For example, 23andMe has penned dozens of partnerships with drug companies for access to parts or all of their collection, putting the company's valuation in excess of $2 billion. In 2012, the Icelandic genomics company deCode was acquired for more than $400 million, which gave Amgen access to a database of more than 100,000 complete or partial (imputed) genome sequences.

Aside from the question of access, another issue with amassing genome sequences or any personally sensitive data in a large, centralized database is that it provides a single entry point for both usage—which puts power and potentially its abuse in the hands of a single person or entity—as well as an entry point for failure; one bad actor could breach the privacy of millions of individuals in one fell swoop. In 2018, for example, MyHeritage's 92 million user accounts were hacked, although only e-mail addresses were obtained, and no DNA data were compromised[2]. And, although not a hack per se, in 2018, the UK National Health Service was found to have passed identifiable medical information from more than 1 million patients on to Google DeepMind to create an app to identify kidney failure, without explicit permission[3,4].

These flaws were on the mind of University of California, Santa Cruz (UCSC) bioinformatics guru David Haussler. In addition to being the Scientific Director of the UCSC Genomics Institute, Haussler is a founder of the Global Alliance for Genomics and Health, a non-profit consortium of more than 500 organizations worldwide that formed in 2013 to create a system for sharing genomic and other sensitive health information. Haussler says that he realized that their vision of having a central repository for the entire world's DNA was never going to happen. "Who would run it? The US Government? No. Google? No. Name an institution that would be globally trusted to run that. There is none, and there won't be any time soon," he says.

In 2014, Haussler took notice of blockchain technology, a nascent cryptography approach that was gaining traction in the financial sector as a means to create an off-market currency marketplace that was immutable, transparent and therefore 'trustless' (Box 1). He floated the idea that blockchain technology might provide a way to democratize data sharing, because it creates a transparent contract within a virtual environment that can automatically perform the functions and checks of transactions. "Patients or advocates could put the results of genetic tests [or] cancer symptoms anonymously on the block chain. Researchers can pull down that information and do research. Even while you're remaining anonymous, you could get research back," he says.

Together with UCSC internet technology specialist and Chief Technology Officer of the Genomics Institute Rob Currie, Haussler built a blockchain, which is currently being used in a pilot program run out of the University of California San Francisco (UCSF), called the Cancer Gene Trust (CGT),

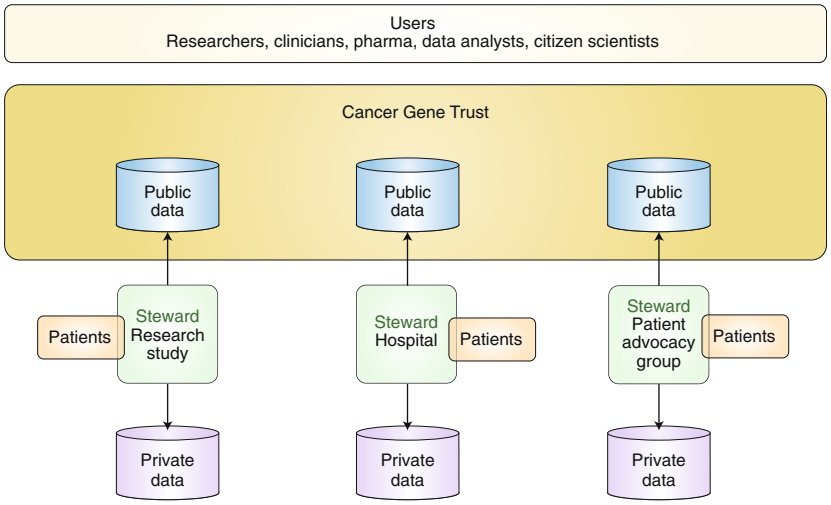

Fig. 1 | Schematic of the Cancer Gene Trust's blockchain architecture. A global off-blockchain decentralized network is controlled by 'stewards' who make various kinds of patient data publicly available. Protected patient data are held only by the steward. If a steward wants to enable contact, a link to a patient's identity can be maintained that is known only to the local steward. Adapted from The Cancer Gene Trust.

to do just that. The system is designed to be a decentralized network that can connect different data sources though a set of nodes, each of which has a steward who retains the data locally and decides what data will be made public and to whom (Fig. 1). The data are then de-identified and submitted to the blockchain. In that way, potential users (researchers) can look on the blockchain and see what is available. In the first use case at UCSF, the researchers went through a clinical trial process and an institutional review board and decided to provide somatic-mutation data and computed tomography scans from community and university hospitals. According to project leader oncologist Eric Collisson of UCSF, "We are trying to put what we think are the critical ingredients for adjudicating clinical trials: what mutation did you have, what drug did you get, what happened to your pixels. We think those are the raw ingredients."

In their pilot program, the researchers have uploaded consented public data (de-identified clinical, imaging and somatic genomic data) to an off-chain data storage site, the inter-planetary file system (IPFS), a peer-to-peer file-sharing system from which large datasets can be distributed and shared—and then submitted the data to the CGT contract. "The primary notable things from CGT are that we captured data generated as part of the standard course of care and actually shared real data via the blockchain, all open source and free, with the intent to show it can be done and provide a starting point for others," says Currie.

Show me the money

Although Haussler and colleagues chose not to incorporate a financial incentive into their blockchain platform, believing that helping others is enough of an incentive, several groups are following that route, creating unique tokens (of dubious value), using existing cryptocurrencies or trading data for genomic services, all of which are meant to incentivize sharing personal health information. The rationale, as articulated by nearly all company spokespeople whom *Nature Biotechnology* spoke with, is this: why allow for-profit companies such as 23andMe (whose 2018 deal with GSK worth $300 million seems to have animated much of this activity) to profit from selling your data, when you can directly pass your data onto pharmaceutical companies or researchers and earn something for it yourself. Blockchain enables this sharing, the data brokers argue, and puts the provenance in data owners' own hands, all the while keeping individual identities private by storing only encrypted (private) keys, instead of personal identifiable information, on the chain (Box 1).

This is where the profit-motive-driven blockchain folk part company with Haussler and his colleagues. "This is tantamount to saying, look, if you give your data out to the biomedical field, they might make a great biomedical discovery for mankind, but you won't get any money out of it. Isn't that horrible? What kind of thinking is that?" Haussler asks.

Be that as it may, several companies working on a DNA marketplace already have platforms that are collecting genomic data along with health surveys or profiles—both kinds of data are needed to create value for potential data buyers. EncrypGen—co-founded by David Koepsell, an ethicist and technology professor, and Vanessa Gonzalez, a genomics researcher at the Universidad Nacional Autónoma de México—was an early adopter of blockchain. Currently, EncrypGen has the only functioning marketplace, which since November 2018 has been brokering financial exchanges between data sellers and data buyers, on something they call the Gene-Chain. Leading up to this, the company issued an initial coin offering (ICO) in 2017, using a token variously called DNA, $DNA and more recently MDNA (in which M presumably stands for money), which garnered the company $1 million. The coin is trading on several crypto-exchanges, exhibiting the typical highs and lows of any currency (albeit a completely untested currency). Through Gene-Chain, individuals can earn MDNA tokens by selling their data, and the tokens can then be stored in virtual wallets, exchanged for services, converted to other cryptocurrencies and someday, according to Koepsell, cashed out.

Also working with blockchain technology is Nebula Genomics, which has George Church among its founders—the same George Church who started Knome, a dot com that offered sequencing services back in 2007, when human genome sequencing cost several thousand dollars, and the Personal Genome Project in 2005, which created a completely open-access platform for sharing personal health data, including personal genome sequence data. Nebula Genomics grants tokens to people who upload phenotypic data to their blockchain. After data owners garner enough tokens, depending on the quantity and type of data they upload, they can purchase their whole-genome sequence. Alternatively, they can either pay out of pocket for their sequencing to be done by Veritas Technologies, a Church-founded company partner of Nebula Genomics or be subsidized by a pharmaceutical company seeking data from people with particular health profiles. The goal, according to Nebula Genomics co-founder and CSO Dennis Grishin, is to make sequencing available universally, not just for the rich, and via the blockchain, to create a secure environment for storing their data and potentially sharing it.

LunaDNA, a non-profit arm of Luna Public Benefit Corporation that was founded by several former Illumina executives, has eschewed cryptocurrencies and blockchain and is instead collecting

Table 1 | Selected companies offering a platform for sharing personal genomic data

Company (founded)	Platform	Currency	Funding	Services	Partnerships
Digital DNAtix (2018)	DNAtix genetic vault, DNAtix distributed genetic storage—supporting Ethereum and Hyperledger blockchains	Internal Tokenization embedded into the platform, on the basis of the ERC-20 framework	Private investors	A digital genetics secured platform for B2B players with connection to a worldwide distributed marketplace of genetic service providers	Feragen, MapMyGenome, Biologix, Morris Kahn Maccabi Health Data Science Institute, more undisclosed
Embleema (2017)	Decentralized permissioned Ethereum blockchain made of HIPAA-compliant nodes allowing patients to share RWE with pharmaceutical companies and health regulators	RWE token	Seed investors $3.7M	Dynamic patient registries, observational studies, safety and efficacy monitoring, clinical-trial optimization	Cystic fibrosis advocacy group, prostate cancer advocacy group, Servier, Pierre Fabre Medicament, IEEE, Republic of Armenia, Beth Israel and others
EncrypGen (2017)	Custom blockchain (GeneChain) for recording transactions/HIPPA-certified cloud storage of de-identified raw data (genotype or WGS)	$DNA token	$1.5M seed round, $1M token sale, investor funding pending	DNA sequencing (through partners)	Microsoft Start-up, Genomics Personalized Health, TPA Network, Codigo 46, Viazoi), Murrieta Genomics), Health Wizz
genomes.io (2017)	Blockchain for private storage and querying of WGS using Secure Encrypted Virtualization (SEV) and Ethereum blockchain	GENE token and Fiat	Seed round of $225K, currently in an investment round	Financial return for allowing selected access to WGS combined with ability to query one's own WGS; secure access and repeat consent mechanisms	Consenys, AMD, TenX Health
LunaDNA/LunaPBC (2017)	HIPPA- and GDPR-compliant storage of de-identified data	Issues (nontransferable) shares	$7.6M seed	Dividend earnings if and when data are used	Genetic Alliance, Awakens
Nebula Genomics (2018)	Exonum blockchain, distributed access control to data, data storage in Google file storage system	Credits redeemable for services	$4.3M Koshla, Arch, F-Prime, Hikma, Mayfield, Mirae, Windham and others	DNA sequencing (through Veritas)	EMD Serono, Veritas
Shivom (2017)	Decentralized permissioned Ethereum blockchain made of HIPAA-compliant nodes	OMX tokens (Ethereum)	ICO presale $35M (28K ETH)	Personalized reporting, DNA kits, DNA data search and marketplace	Living DNA, VItl, Chronomics, Family Care Path, Lympho, Lifebit
Zenome (2017)	Ethereum distributed block chain, smart contracts	ZNA Ethereum 35M	$100K private investors, $200K token presale (2017), $360K ICO (2018)[a]	Sequence services (WGS or exome) for $200–500	BGI, Helicon), SberX, Genetico, Skkoltech
WuXI NextCODE	Permissioned Ethereum-based blockchain (LifeCODE.ai), anonymized encrypted decentralized data storage	LifeCODE (built-in) token (LCT)	$440M (not all related to LifeCODE)	LaiyinTribe app	WeGene (DTC genomics)

WGS, whole-genome sequencing; M, million; K, thousand; FHIR, fast healthcare interoperability resources; CCD, continuity of care document; FDA, US Food and Drug Administration; IEEE, Institute of Electrical and Electronics Engineers; ETH, Ethereum; B2B, business to business; D2C, direct to consumer. [a]Values of cryptocurrencies fluctuate over time.

data into a secure cloud-based platform. Instead of cryptocurrency, Luna grants shares of the company to people in exchange for providing data—the number of shares granted depends on the precise data provided and is laid out in their filing with the US Securities and Exchange Commission[5]—making users in essence part owners of the company and hence eligible for dividends once the information aggregated by LunaDNA has value in the marketplace. LunaDNA President Dawn Barry calls what they are doing creating a community of sharers, and in this community, data are the currency.

At Embleema, which was founded by several individuals with experience in large-data collection, chief scientist Vahan Simonyan developed HIVE, a parallel distributed computing environment that he says is "tailor made to put in the marketplace." HIVE has a blockchain with more than 50,000 patient records. Data owners earn what they call real-world evidence (RWE) tokens by uploading data, and data purchasers buy RWE tokens from the data owners by using hard currencies to obtain access to the data. Currently, Embleema charges a small management fee for indexing and validating the data, but in the next phase, they will start engaging with the marketplace, which will generate value for the data owners.

Show me the data

Data collection is where these enterprises start, which is relatively straightforward because it has been going on for decades in both public and commercial databases. Whereas some companies provide sequencing services if needed (Table 1), much of what is being uploaded is genotypic marker data (not whole-genome sequence) initially obtained from third parties such as 23andMe and Ancestry. Koepsell says that approximately 50% of the data on the Gene-Chain are genotypes from 23andMe. The consumer genetics giant, as well as several other major DTC genomics companies—MyHeritage, Family Tree DNA, National Geographic's Geno and Ancestry—enable their customers to download their raw data (single-nucleotide-polymorphism chip data for hundreds of thousands of alleles of significance) and do whatever they wish with it.

But to be truly useful to researchers, these companies would need thousands, if not hundreds of thousands, of genotypes. Consequently, we are starting to see the companies in this group forging partnerships with patient-advocacy groups and foundations, health institutions and even governments to get large numbers of people onto their platform. LunaDNA, for example, has a partnership with Genetic Alliance, which, through the integration of the PEER platform, will give 50,000 patients across 45 disease communities access to the benefits of belonging to the LunaDNA community. Sharon Terry, President and CEO of Genetic Alliance, says that they had been trying to think of how to ethically compensate individuals for sharing their data. "This is a nice balance. It's not tons of money. We're not paying people for data; we're not buying data or selling data. We're compensating people for sharing data and people actually benefit from research activity, which they should." The London-based blockchain company Shivom is working with several governments—the government of Andhra Pradesh, India and the government of Malta—on potentially particular disease indications that are rampant among the relevant populations. EncrypGen recently announced a partnership with third-party administrators, the TPA Network, a consortium of entities that process insurance claims for self-funded communities. Through this partnership, EncrypGen could potentially gain access to health data from over 100,000 individuals covered through TPA Network. In addition, Nebula Genomics recently announced an arrangement with EMD Serono in a project directed at enticing individuals with lung cancer to join their platform. Serono will provide free whole-genome sequencing to qualifying lung cancer patients in exchange for access to the Nebula Genomics platform (Nat. Biotechnol. **37**, 706, 2019).

Where's the beef

Blockchain is tailor made to set up and memorialize transactions. As Haussler puts it, it contains "everything you would want in a contract: execute only in the correct order, with the correct conditions met." He continues, "Instead of having to put your trust in a third party or a specific government, you can put your trust in this other agency, especially when you are making international transactions."

However, blockchain is not suited for data storage when large amounts of data are involved, especially given that with genomic data rather than currency transactions, there is an entity—the data—that must be both made available and protected, capabilities that blockchain by itself cannot provide. Thus, other systems for data storage and distribution must be bootstrapped onto the blockchain. Bradley Malin, bioinformatics expert at Vanderbilt and co-chair of the security working group of the US National Institutes of Health (NIH) All of Us Research Program, explains, "Blockchain was designed with currency-based transactions in mind and a way to prevent fraud in such space. What we're seeing now is that people are attempting to build around it, so that you can achieve other properties, such as confidentiality and security, but in general it has been quite challenging to realize this, and particularly at scale."

In most instances in which blockchain is used for personal health data, the data are held either on an individual's own computer or more likely are placed in a commercial cloud with Health Information Portability and Accountability Act (HIPPA)-compliant security in place, and, once a transaction is agreed upon, the data are made accessible to the user. How that process occurs differs among companies. Some hand over anonymized and encrypted data to buyers, whereas others set up environments (containers) in which the actual analysis is performed, which means that the data do not move, and the computational pipeline comes to the data. Technology development for this process has been ongoing, because the problem of sharing large datasets exists with all types of data and their uses, and it is not specific to blockchain platforms. As Currie describes, "We're at a point where the general, big-data genomics community is moving to packaging their stuff and running it at arm's length because they have to; they can't move the data."

What could possibly go wrong?

The need for more personal genome and health data to galvanize research on chronic, complex diseases is indisputable. Whether blockchain is the solution to accelerate that process is unclear. MacArthur agrees that 'siloing' is a real problem and technology development is needed, and thinks it is good that people are experimenting with different models. "We are at a point in human history where we know that an enormous amount of genetic data will be generated over the next few years, so there is a real urgency to develop a good model for storing, protecting and making that data accessible to researchers," he says.

Whether the profit motive will be a help or a hindrance to achieving the goal remains to be seen. These efforts at making personal genomic data accessible may well have the opposite effect in just creating more silos. But Vanderbilt's Malin doesn't see them as competing in the same space with, for example, the NIH's All of Us Research Program. "I don't think the people who put their data on a blockchain environment would be the ones participating in programs like All of Us to begin with. People on blockchain are already technologically savvy, and they have other motives that are driving them."

UCSF's Collisson also thinks the space needs shaking up: "I think a lot of medical centers are sitting on the side lines, being like, 'Well we can't sell the data today, but we think it would be great if someone gave us a billion dollars for them. If we give them away for free, then we certainly won't get any money for it. Let's do nothing.'"

Laura Hercher, a genetic counselor and Director of Research at Sarah Lawrence College's Program in Human Genetics, finds the language being used to describe these new endeavors "high blown." She continues, "Liberation of your genetic information feels more like a movement rather than an industry." Yet she points out that "the companies are part of an industry, and ultimately they have to make a profit."

Whether the commercial blockchain will take off remains to be seen. It will work only if a sufficient critical mass of individuals enroll in the services. There is currently little information available on who is engaging with companies in the personal health data marketplace.

Martijn van Kalsbeek may be a prototypical early adopter of the approach. He's tech savvy and idealistic, and he was

trading in cryptocurrency for a few years before the Gene-Chain service became available. He joined because he anticipates making a profit from his data (which he already has to a small degree) as well as providing them to the scientific community anonymously to galvanize biomedical research. He expects his tokens to increase in value as the marketplace grows, but, of course, that remains to be seen.

In the meantime, he says that he derives satisfaction from the simple act of donating his data. And the Gene-Chain, he says, puts him in control of who has his data and more pointedly who controls "the sale of [his] data." ❐

Laura DeFrancesco[1]* and Ariel Klevecz[2]
[1]*Pasadena, California, USA.* [2]*Santa Fe, New Mexico, USA.*
**e-mail: l.defrancesco@us.nature.com*

Published online: 16 July 2019
https://doi.org/10.1038/s41587-019-0200-5

References

1. Regalado, A. *MIT Technology Review* https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/ (11 February 2019).
2. *MyHeritage Blog* https://blog.myheritage.com/2018/06/myheritage-statement-about-a-cybersecurity-incident/ (4 June, 2018).
3. Hill, R., *The Register* https://www.theregister.co.uk/2018/06/13/royal_free_deepmind_deal_audit/ (13 June 2018).
4. Lomas, N. *TechCrunch* https://techcrunch.com/2016/05/04/concerns-raised-over-broad-scope-of-deepmind-nhs-health-data-sharing-deal/ (4 May 2016).
5. *U.S. Securities and Exchange Commission* https://www.sec.gov/Archives/edgar/data/1741687/000149315218014113/partiiandiii.htm (4 October 2018).