



# Blockchain enabled genome security from the moment it is sequenced

The genomes.io Whitepaper V 1.1.4

Dr Mark Hahnel

The genomes.io Whitepaper V 1.1.4	1
<b>Abstract</b>	<b>3</b>
<b>The Market</b>	<b>4</b>
<b>Genomic Blockchain Consortium</b>	<b>5</b>
<b>Competitors in the Space</b>	<b>5</b>
<b>Use Cases</b>	<b>6</b>
Genomes Store	6
The Personal Genomes Marketplace	7
Individuals	8
Research Groups	8
Pharmaceutical Companies	9
<b>The Technology</b>	<b>9</b>
Summary	9
Technical Insights	10
Rockchain	10
PHE-BLOOM	11
<b>The GENE token</b>	<b>11</b>
<b>Revenue Model &amp; Use of GENE Tokens</b>	<b>11</b>
Controlled Queries	11
Examples of questions - End user	11
Examples of questions - Industry	12
<b>The Team</b>	<b>12</b>
<b>Roadmap</b>	<b>14</b>
<b>References</b>	<b>14</b>
Example Query Script	16

## Abstract

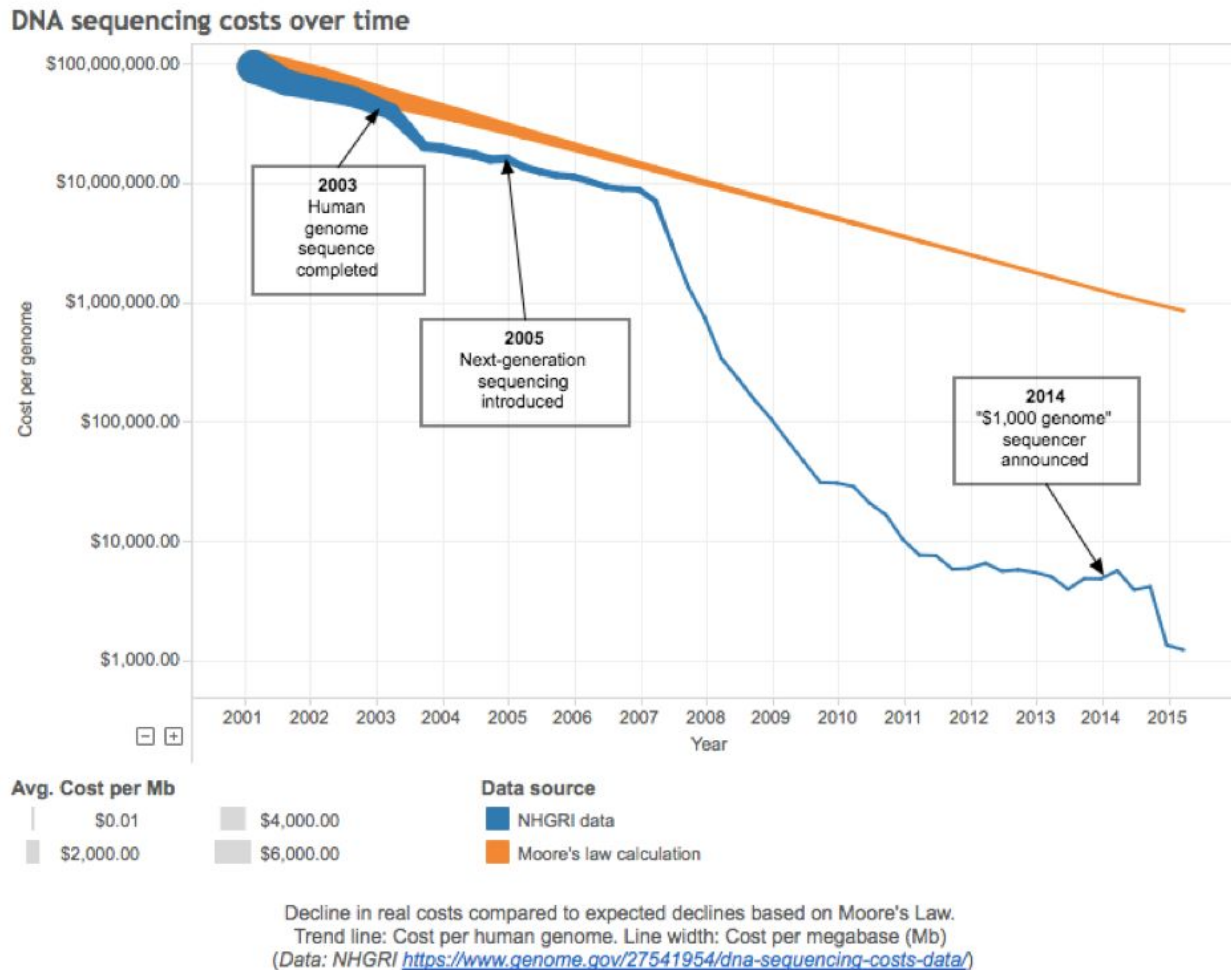
Sequencing of human genomes is fast becoming an affordable technology for general consumers. The huge drop in operational costs for this technology, combined with the enormous potential for personalised medicine, means that we are looking at a future where all of humanity, all 7B+ people have their genomes sequenced (Cotton 2009; Zwart 2009). The potential for exploitation of people's personal genomic data is therefore a huge concern (White 1999). Securely storing genomes in a query-able way is an unrealised necessity that is now realisable.

'Genomes' is a blockchain application designed to privately, securely store whole genome sequence data for an individual in a manner that allows questions to be asked of the genetic code without exposing the code to others. Genomes.io uses Ethereum and Rockchain to allow systems to compute a function on private data, without exposing anything about their data besides the result. All corporate data are held locally and securely, the permissions are finely tuned through Ethereum blockchain smart contracts. The only data transferred outside the company is the computation results. The ComputeReduce node can also serve as a connection proxy from the corporate world to the outside world, such as traditional firewalls or internet proxy servers.

'Genomes' not only secures privacy for an individual's genomic data, it also enables 3<sup>rd</sup> party access when approved by the individual in exchange for tokens known as GENE. These GENE tokens can be sold on the free market, back to the genomes organisation, or to other users through the app. These tokens (GENE) can also be used by pharmaceutical companies to reward users of the app, who will expose certain computational results to 3<sup>rd</sup> parties. Genomes.io will make use of the [human phenotype ontology](#) to interoperate with leading biomedical research (Smedley 2017).

# The Market

Fig 1. DNA sequencing costs over time



In 2001, after 15 years of work and a bill of \$2.7B, the Human Genome Project successfully sequenced the first human genome. As of 2017, the cost of sequencing the human genome has dropped to under \$1000.

*"The '\$1,000 genome' has become shorthand for the promise of DNA-sequencing capability made so affordable that individuals might think the once-in-a-lifetime expenditure to have a full personal genome sequence read to a disk for doctors to reference is worthwhile"* ("\$1,000 Genome - Wikipedia" n.d.).

With affordable costs combined with a stronger toolset for genome editing. Eg. CRISPR, the demand for human genome sequencing is set to grow exponentially to the point where all of humanity will see their genome as akin to their birth certificate. This raises enormous privacy

and exploitation concerns globally (Goodman 2016). In order to responsibly speed up genetic diagnostics and realise the potential of personalised medicine, a workflow must exist whereby individuals have complete control over their genome and access to it, from the point it is sequenced.

'Genomes' is set up with the aim to secure this workflow in a manner that is only now achievable. 'Genomes' aims to sequence 1 billion genomes and provide secure services on top of them. We endeavour to partner with existing global efforts and collaborate with hardware providers to drive costs down further and make the personalised human genome a reality for all of mankind within a generation.

Individuals, research groups and pharmaceutical companies will be able to query their own and request to query other genomes anonymously. Each transaction will use tokens called GENE in order to be processed.

## Genomic Blockchain Consortium

























Genomes.io is a member of the Genomic Blockchain Consortium, which is being set up with EncrypGen and MyGenomeBank to address the following concerns:

1. Form a Consortium of Genomic Blockchains
2. Organize as a separate not-for-profit entity this consortium
3. Create a democratic governance structure
4. Under the consortium, propose, vote on, and establish technical as well as ethical standards for the purposes of:
  - a. ensuring that consumers and their data are portable, safely preserved, and protected
  - b. creating some interoperability at some level to better serve the searching for, payment, and use of genomic data by platform customers
  - c. devising ethical guidelines to protect privacy of user data across platforms

We believe that we have a great opportunity now, before the wide adoption of any particular genomic blockchain, to create an environment that both fosters competition and helps advance the goals of consumer empowerment and scientific advancements, that we have all stated publicly, motivate our products and services .

# Competitors in the Space

Table 1. Comparison of genomic sequencing companies with a focus on blockchain

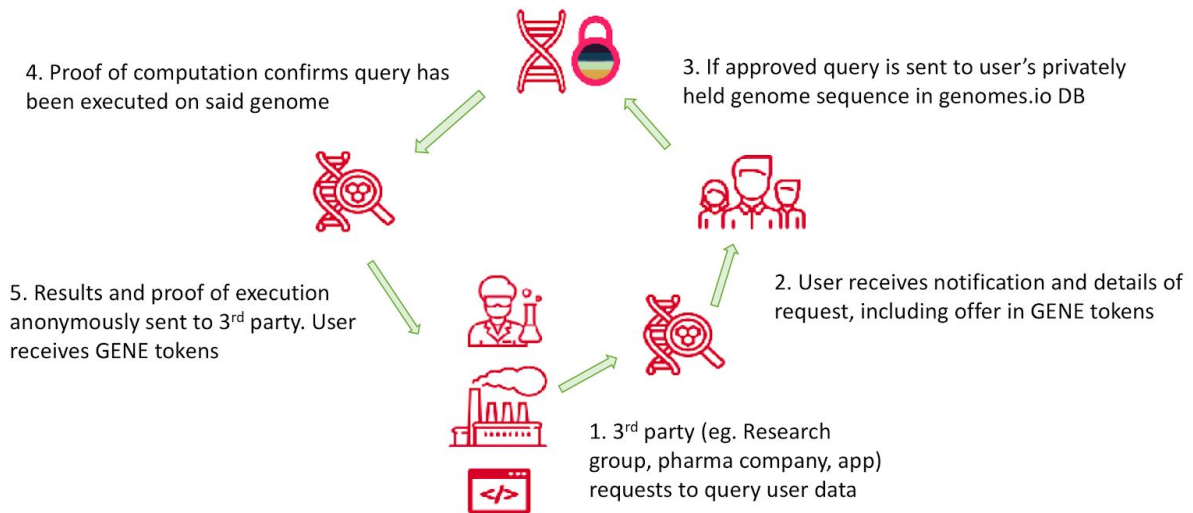
					
Full Data Anonymity					
Genomic Sequencing Marketplace					
DNA Marketplace					
Support of 3rd Party Apps					

## Use Cases

### The Genomes Store

The Genomes store will allow apps and packages to be built on top of the Genomes' query layer. This will allow app developers to construct tools that can provide insights into an individual's genetic make-up, from ancestry to fitness, health and nutrition. The app makers can charge a premium for individuals to make use of the apps. These premiums will be paid in the GENE token.

## Genomes Store



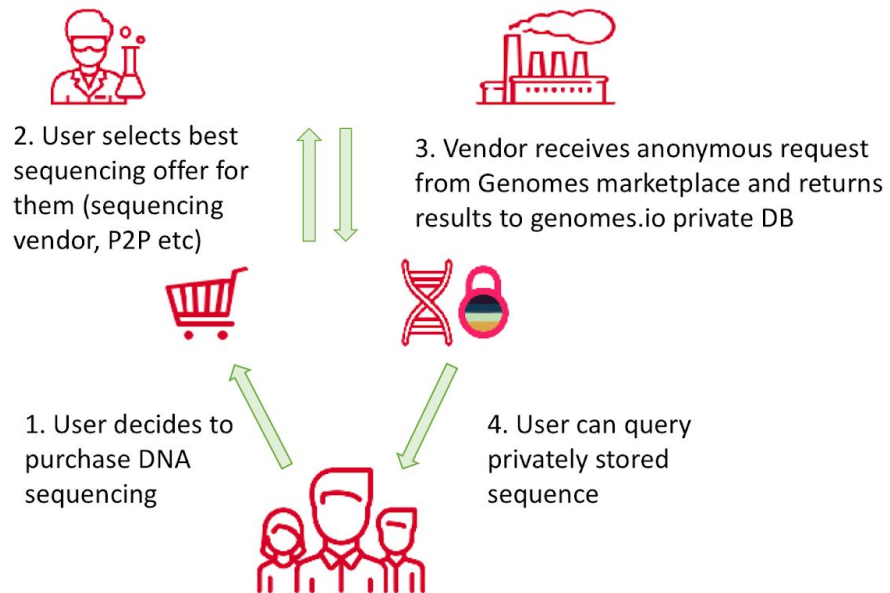
## The Personal Genomes Marketplace

As the need for genomic sequencing increases, the idea of companies, start-ups and even individuals competing on the price offered for such services will drive prices lower. New technologies, such as Nanopore mean that the idea of anonymised peer to peer genomic sequencing is on the horizon. The Personal Genomes Marketplace, will allow end users to choose who sequences their genome, based on a range of criteria, such as:

- The cheapest
- The best applications in the genomes store included in the package
- The most established
- A company that offers discounts in exchange for privately querying a certain phenotype

All transactions will be powered anonymously using the GENE token.

## Personal Genomes Marketplace



## Individuals

Any patient or customer can put their DNA file on a local Rocker instance and cherry pick the datascripts (rules) they want to be applied on their files. Those datascripts can offer advice about food consumption, or inform about disease risks. Optionally, the user could accept execution of datascripts from third party providers. We see factors such as family planning, drug dosage prescription and diagnosis of genetic diseases as being drivers in the future of personalised genomics.

An app for users to intelligently query their own genome, in the way that 23andMe allows is being developed. This will process the data dynamically upon loading, based on queries applied through the Rockchain framework. This means that no one else at Genomes.io or anywhere else, will have access to the data.

## Research Groups

Academic research groups looking to investigate particular genetic pathways often do not have access to large amounts of human genomic sequences. This has led to several logistically difficult projects, such as the [100,000 genomes project](#). As a result of this, there are a wide range of existing high-quality tools to analyse such data, but no data to do so. Allowing requests from research groups will be a much more efficient use of public and private grant funding, in an



ethical manner for all involved. Decentralization is the key to build human genomic networks in this manner.

## Pharmaceutical Companies

Pharmaceutical companies have previously spent millions of dollars attempting to obtain large numbers of human genetic data to query. The most recent example of this is Genentech investing \$60m in 23andMe in order to obtain access to the ~1,000,000 exomes that have been sequenced. This also represents how commercial players in the genetic sequencing space are selling individuals data in order to sustain their profitability. This can only increase as more people get their genomes sequenced.

By allowing the pharmaceuticals to request permissions to query individual's data with no middle man in a transparent manner, the pharmaceutical industry achieves a more scalable version of their goals at a much reduced cost and the individuals maintain as much privacy of their genetic code as they wish.

Pharmaceutical companies can pre-buy GENE tokens today, by enquiring about partnerships at [info@genomes.io](mailto:info@genomes.io)

## The Technology

Real private querying of genomic data:

Based on 'Proof of documents transfers between nodes' and 'Private node to node computation'

## Summary

Rockchain is a distributed data intelligence platform that focuses on access rights orchestration. It can define how files are exchanged between peers, which parts of the files can be exchanged and what computation can be made on said files. This "Extreme Privacy" Engine is only about data: all data exchanges, all transactions are kept on the Ethereum blockchain (they are "notarized"). Applied to genomes, Rockchain can perform MapReduce operations on a distributed set of distinct genomes to perform basic genetic querying, without compromising any of the genome privacy.

## Technical Insights

### Rockchain

Rockchain is a distributed data intelligence platform based on an asynchronous message-passing model between nodes. Rockchain has two types of nodes each one having a specific role in the mapreduce algorithm. The fact map nodes contain the data and filters the data, based on their structure. They can also define data channels between fact map nodes for automatic read/write synchronization. Fact map nodes never communicate data to reduce nodes, except if they are asked to do so in the data processing rule. The reduce nodes are executing data reduction logic, often pairing their logic with some of the fact map nodes using advanced cryptographic techniques. Reduce nodes are executing Javascript logic collectively, i.e. like in Ethereum, several reduce nodes are executing the same logic and are comparing the javascript execution results.

Rockchain javascript restricted language (internally named datascript) secures the execution through merklized abstract syntax trees, providing a secure cryptographic proof of execution, a proof of code integrity, as well as a proof of code execution multiplicity on several nodes. Each computation step is guaranteed by a unique hash and all the above proofs are documented on the Ethereum Blockchain. Rockchain uses a customized version of the docker container, called Rocker, that extends the Docker 1.10 cryptographic proof of docker machine layers to datascript deployment, ensuring also the proof of version for all deployed datascripts. It is also extending the native Docker networking layer used in Docker compose to add peer-to-peer dynamic networking features among reduce nodes belonging to the same cluster.

Rockchain thus extends the Docker ease of deployment facilities to a global datascript repository, maintained in a decentralized DataScript Name Service. It is permanently storing datascript execution proofs, as well as computation results proofs in the Ethereum Blockchain. It provides a proof of process for process traceability, process compliance, while maintaining data privacy protection. It also provides real time accountability for all data processing tasks on all distributed processes.

All fact map nodes keep their data private whereas reduce nodes can have their data either; public, or restricted to all nodes dealing with a specific rule. In case of reduce data computation on private data, advanced secured two-party computation protocols allow reduce nodes to perform computation while keeping the data on fact map nodes. All computations performed on the fact map and reduce nodes are cryptographically verifiable computations on cryptographically verifiable algorithm. Cryptographic proofs of executed code on provable deployed datascripts can open up truly secured and provable computation cloud services platform, which have not yet been available in the cryptosphere.

## PHE-BLOOM

Genomes.io will make use of the Partially homomorphic encryption (PHE)-Bloom approach for genetic disease testing using homomorphically encrypted Bloom filters. PHE bloom is a modification of FHE-BLOOM, that uses Fully homomorphic encryption (FHE) which enables the cloud to compute a match under encryption by multiplying query and patients' Bloom filters and to aggregate the results into single ciphertexts that are returned to the data owner. Note that this scheme slightly leaks access patterns, e.g., the cloud may learn when a query is posed twice. However, the actual contents of a query and, importantly, the patients' data are still fully protected. FHE-BLOOM performs a disease susceptibility test on a database with 50 patients with up to 100000 variations in approximately 5 min. PHE-BLOOM notably decreases this by four orders of magnitude to 75 ms (Ziegeldorf et al. 2017).

## The GENE token

Total GENE supply - 200 million

100 million to be allocated to investors

100 million held for industry purchase

Full sales price 1 ETH = 10000 GENE

## Revenue Model & Use of GENE Tokens

### Controlled Queries

Like Rockchain and Ethereum, Genomes.io earns fees for each data intelligence module call. Each question a user asks of their sequences will require GENE which acts as transaction costs. The GENE can be purchased in app or on token trading exchanges.

Examples of questions - End user

*Do I have an elevated risk of Breast Cancer?*

*What is my genetic ancestry breakdown?*

*Am I allergic to x?*

Users will be able to ask their own questions. Genomes.io will translate natural language questions into calls to Rockchain datascripts (which will analyse the genome characteristics). Each natural language question will have a price in GENE tokens.

Upon establishing the genomes app ecosystem, large scale analysis of Data can be carried out and users who have had their DNA sequenced can sell access to their information to 3<sup>rd</sup> parties wishing to ask specific questions of their information. The users can agree to this on a case by case basis.

Examples of questions - Industry

*What percentage of the population express this gene?*

In this scenario, users will receive a notification that informs them:

*Pharmaceutical company x would like to query your data to see if you express the gene XXX, they will not receive any other personally identifiable data. They are offering 2 GENE in exchange for this information. Approve? Yes/No*

This workflow is unique in rewarding users for allowing anonymous querying of their data. The pharmaceutical company has access to specific information in an ethical and mutually rewarding relationship.

## The Team

Dr Mark Hahnel

Mark obtained a PhD in Stem cell biology from Imperial College in 2012 after previously studying Genetics and Human Genetics at the University of Newcastle and Leeds respectively. Mark is the current CEO of figshare, a SaaS platform that has provided data infrastructure for the world's leading academic publishers (SpringerNature, Wiley, PLOS, ACS), academic institutions, academic funders and pharmaceutical companies.

Dr Manuel Corpas

Dr. Corpas is an experienced researcher, trainer, and scientific communications strategist. He was the Scientific Lead at Repositive, a genome data sharing platform company. Before Repositive, Corpas was a 2016 fellow of the Software Sustainability Institute and Project Leader at The Genome Analysis Centre (now Earlham Institute, Norwich, UK). He holds a Ph.D. in bioinformatics from the University of Manchester, UK, and did his postdoctoral work in clinical genomics at the Wellcome Trust Sanger Institute, Cambridge (UK).

Sébastien Jehan

Sebastien is a serial entrepreneur with a strong experience in building complex international IT systems (mostly in banks). He is a Telecom Engineer graduate (TelecomParisSud), with a MSc in Finance (ICMA), and has focused his career on Ethereum since 2015. His latest project founded from scratch was Ledgys.io. Sebastien is also an advisor on Sangus.org, the first distributed investment research network. Sebastien has experienced several cultural work environment (US, Japan, Europe) and has experience in Devops automation and Datascience (he's also animating a meetup Datascience in Finance).

Djemba Bocoum

Djeneba is graduating from ETNA Paris (Architecture Systems, Networks and Security), and a graduate in Electronics and Industrial data at Moulay Ismail University in Morocco, in Electronics and Automatics at University of Sciences and Technology of Lille 1. She is a Go developer, working on adapting the RockEngine on smart devices and IOT devices.

Régis Gourdel

Régis has a strong theoretical background in cryptography and is a GO developer. Régis is a graduate from Telecom ParisTech (applied mathematics on cryptography and quantum computing), and is doing an MSc in applicable mathematics in London School of Economics (game theory, advanced algorithms and financial maths).

Simon Redhead

Simon joins us for his project management and community engagement skills, straight out of a role in UK government

Chris Griffiths

Chris is our international business lead, building on top of his background as an engineer and strong global network

Aldo de Pape

Aldo is the founder & CEO of TeachPitch brought onboard to offer business development expertise. TeachPitch is a cloud based platform that helps teachers & schools resolve the problem of information abundance through the process of curation, online tutoring and artificial intelligence. Aldo previously worked in publishing for Springer and Macmillan and is the author of a children's book 'I am!', published in 2008.

## Roadmap

March 2018 – 'Genomes' POC established in the form of a FactMap node network will be with an API.

March 2018 – Pre-sale

Q2 2018 – Work begins on 'Genomes' DAPP on top of FactMap node network (mobile and Web DAPP for desktops) that connects to Rockchain.

Q3 2018 – Initial Coin Offering

Q3 2018 - Recruitment Drive, establishment of 'Genomes Store' and 'Personal Genomes Marketplace'

Q4 2018 – All ICO investors over 10ETH have their genome sequenced

Q2 2019 – Genome sequencing services offered

## References

"\$1,000 Genome - Wikipedia." n.d. Accessed March 3, 2018.

[https://en.wikipedia.org/wiki/\\$1,000\\_genome](https://en.wikipedia.org/wiki/$1,000_genome).

Cotton, Richard G. H. 2009. "Response to Stenson et Al. in Human Genomics Vol. 4, No. 2, Pp. 69-72: 'The Human Gene Mutation Database: Providing a Comprehensive Central Mutation Database for Molecular Diagnostics and Personalised Genomics.'" *Human Genomics* 4 (3): 145.

Goodman, Bryce. 2016. "What's Wrong with the Right to Genetic Privacy: Beyond Exceptionalism, Parochialism and Adventitious Ethics." In *Law, Governance and Technology Series*, 139–67.

Smedley, Damian. 2017. "Faculty of 1000 Evaluation for The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease." *F1000 - Post-Publication Peer Review of the Biomedical Literature*. <https://doi.org/10.3410/f.725602763.793528156>.

White, M. T. 1999. "Underlying Ambiguities in Genetic Privacy Legislation." *Genetic Testing* 3

(4): 341–45.

Ziegeldorf, Jan Henrik, Jan Pennekamp, David Hellmanns, Felix Schwinger, Ike Kunze, Martin Henze, Jens Hiller, Roman Matzutt, and Klaus Wehrle. 2017. “BLOOM: BLoom Filter Based Oblivious Outsourced Matchings.” *BMC Medical Genomics* 10 (Suppl 2): 44.

Zwart, Hub. 2009. “From Utopia to Science: Challenges of Personalised Genomics Information for Health Management and Health Enhancement.” *Medicine Studies* 1 (2): 155–66.