# Genomes – Who gets to read your DNA?

Mark Hahnel



## Abstract

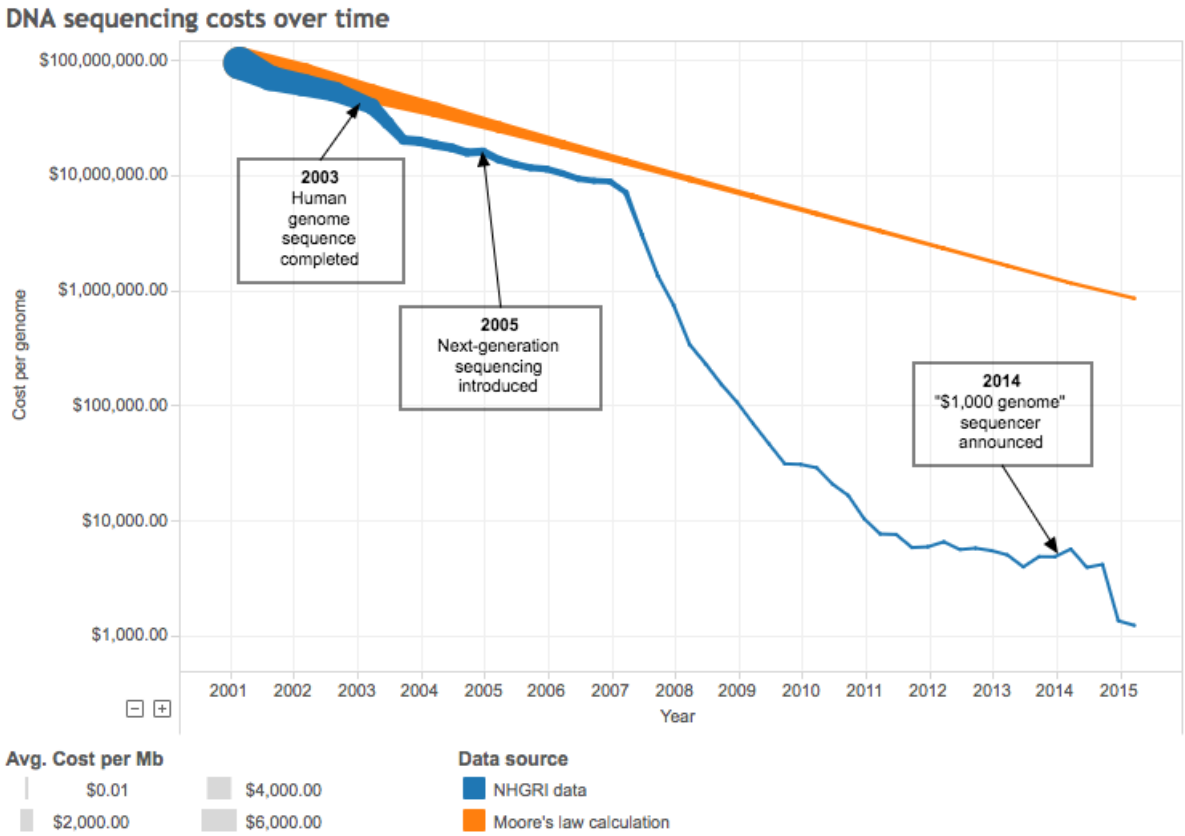
Sequencing of human genomes is fast becoming an affordable technology for general consumers. The huge drop in operational costs for this technology, combined with the enormous potential for personalised medicine, means that we are looking at a future where all of humanity, all 7B+ people have their genomes sequenced. The potential for exploitation of people's personal genomic data is therefore a huge concern. 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 individual's genomic data, it also enables 3rd party access when approved by the individual in exchange for tokens known as OME. These OME tokens can be sold on the free market, back to the genomes organisation, or to other users through the app. These tokens (OME) can also be used by pharmaceutical companies to reward users of the app, who will expose certain computational results to 3rd parties. Genomes.io will make use of the [human phonotype ontology] to interoperate with leading biomedical research.

## The Market




Decline in real costs compared to expected declines based on Moore's Law.
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)
(*Data: NHGRI https://www.genome.gov/27541954/dna-sequencing-costs-data/*)

In 2001, after 15 years of work and a bill of $2.7B, the Human Genome Project sucessfully 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” https://en.wikipedia.org/wiki/$1,000_genome

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. 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 OME in order to be processed.

## Secure DNA consortium (SDC)

In order to encourage fair usage across service providers, genomes.io will also work with leading genome sequencing companies to allow potential access to a larger pool of genetic information.

### Partner sequencers

TBA

### Partner pharmaceutical companies

TBA

## Use Cases

### 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.

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.

## The Technology

### 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 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 secured 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. Rockhain 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.

## Revenue Model & Use of OME Tokens

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

Examples of questions can be:

*Do I have an elevated risk of Breast Cancer?*
*What is my genetic ancestral 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 OME 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 3rd parties wishing to ask specific questions of their information. The users can agree to this on a case by case basis.

Examples of questions can be:

*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 OME 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

Mark Hahnel, PhD.

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.

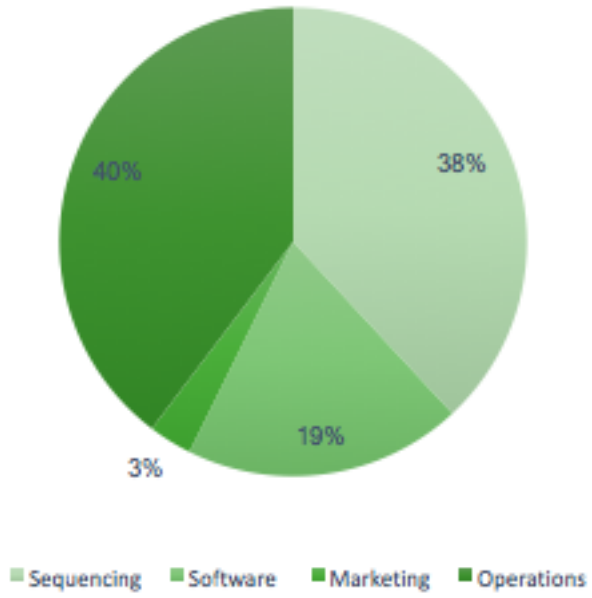
## Roadmap

August 2017 – 'Genomes' POC established in the form of a FactMap node network will be with an API.
Sep – Dec 2017 – Work begins on 'Genomes' DAPP on top of FactMap node network (mobile and Web DAPP for desktops) that connects to Rockchain.

Oct/Nov 2017 - Initial Coin Offering
Q1/Q2 2018 – Recruitment Drive, first sequencing laboratory established in London
Q1-4 2018 – All ICO investors over value 50ETH have their genome sequenced
Q1-4 2018 – 'Genomes' iOS and Android app beta launch
Q1 2019 – Genome sequencing services offered
Q3 2019 – 'Genomes' app ecosystem established

## Costing

### Funding Breakdown




### Sequencing
Genetic Sequencing Infrastructure €6m
Illumina NovaSeq 6000. 2 x €1m
Cost of sequencing for qualifying ICO participants €1.5m
Operations for sequencing lab €1.6m/year for 3 years

### Software
Open source app layer development €600k for 3 years
DAPP development team - €1m/year for 3 years

### Marketing
Marketing €250,000/year for 3 years

### Operations
Legal team - €1m/year for 3 years
Management, business development & Operations - €750/year for 3 years

## ICO

### Target

€25,000,000
110,000,000 OME (10 million reserved for founders)

The ‘Genomes’ Initial Coin Offering will be for 100 million tokens, known as OME. The ‘Genomes’ ICO will issue the 100m OME at a rate of 1 ETH = XXX OME (TBD in week of ICO).

The OME Token is an essential part of the Genomes.io ecosystem and economy.

### Dates

The ICO will run for 1 month from dd/mm/yy to dd/mm/yy.