The Zenome Project: Whitepaper
blockchain-based genomic ecosystem

Nikolay Kulemin Sergey Popov Alexey Gorbachev

October 6, 2017

Abstract

Industry 4.0 is a title for the current trend of automation, scaling and data exchange in manufacturing technologies. It includes artificial intelligence, virtual reality, the Internet of things and Big Data analysis. Genomics is a vivid representative of the industry 4.0 that requires solving many urgent problems, such as storage and analysis of Big Data with keeping public access for researchers and privacy for people.

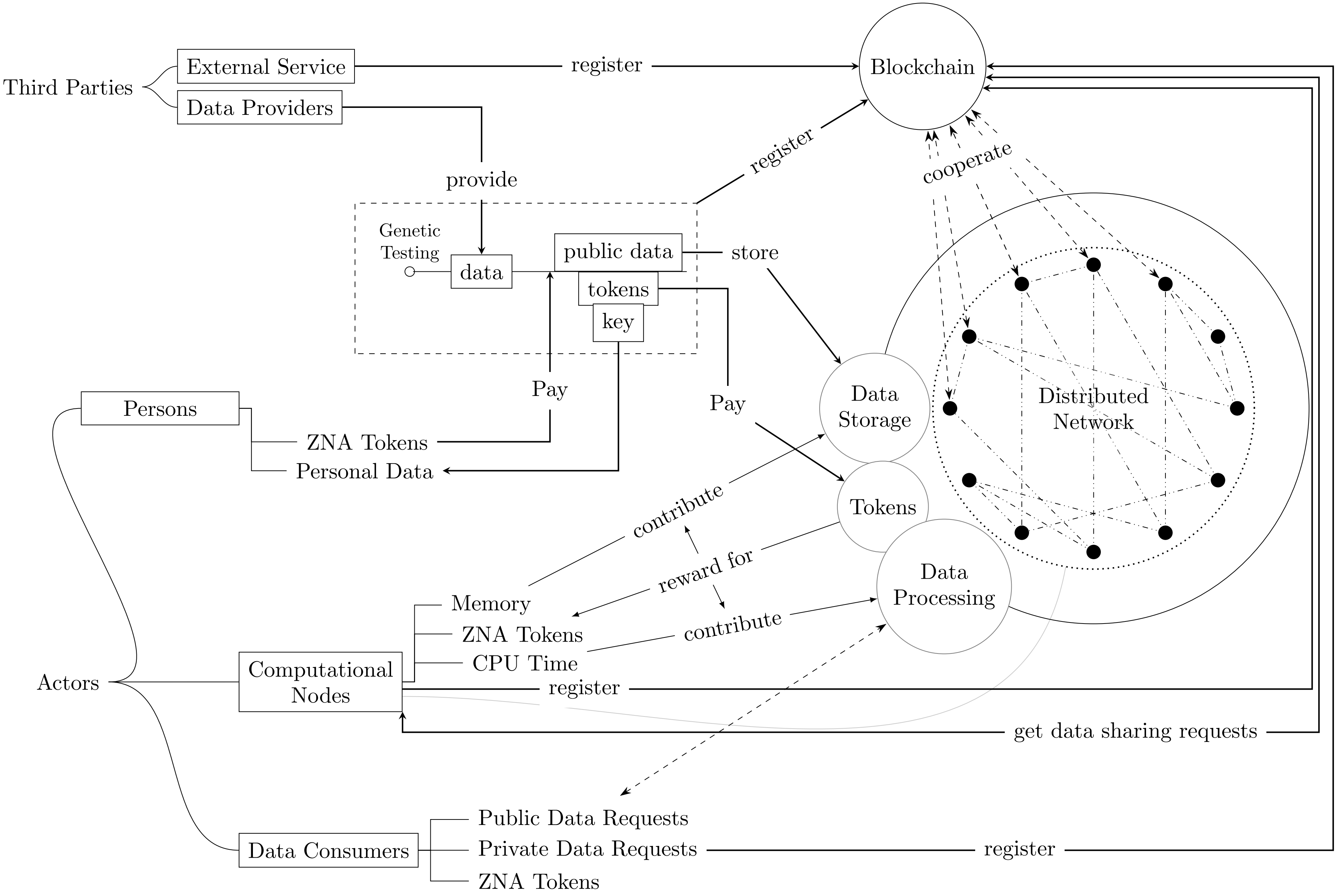
Currently there is a problem of inequality in the genome industry. It means that the main part of personal genomic data was concentrated in data centers of genomic corporations, government, scientific and medical institutions and pharmaceutical companies. Moreover, there is an issue of legal limitation of access to personal genomic data, as well as the absence of possibility for genomic data management and sharing. This genomic data monopolization dramatically inhibits the development in a number of scientific and medical fields.

The development of cryptocurrencies and blockchain-based technologies leads to significant transformation of many economical domains. Application of blockchain approach will be a lifeline allowing to upgrade the development of personal genomics. It will make each person the owner of his or her genetic data.

The Zenome project is a decentralized blockchain-driven database of genomic information. This platform supports the possibility to manage your genomic data while maintaining privacy and ability to make a profit from selling access to different parts of the genome. It will establish equal conditions for drug development and for the progress of scientific and medical technologies.

Zenome is a new economic environment based on genomic data and blockchain technology. The implementation of our conceptual model will solve the following difficulties:

- Creating an infrastructure for storing Big genomic data using distributed database
- Open access to millions of human genomes worldwide with privacy protection
- Possibility for each person to participate in scientific and clinical research and to make profit from this
- Stimulating the enhancement of genomic sciences in developing countries and de-monopolization of genomic data in developed countries


Third Parties
External Service
Data Providers
register
Blockchain
provide
register
cooperate
Genetic Testing
data
public data
store
tokens
key
Pay
Pay
Persons
ZNA Tokens
Personal Data
Data Storage
Tokens
Distributed Network
contribute
reward for
Data Processing
Memory
ZNA Tokens
contribute
CPU Time
Actors
Computational Nodes
register
get data sharing requests
Public Data Requests
Private Data Requests
register
Data Consumers
ZNA Tokens

The Team

Alexey Gorbachev

Founder

Molecular biologist and Blockchain Enthusiast
Ph.D. in molecular biology and biochemistry

University: Moscow State University, Department of Molecular Biology.

Alexey has a significant scientific background, expertise in business, and project management. In Zenome Alexey is responsible for the overall vision and development of the business.

Email: alex@zenome.io
LinkedIn: `https://www.linkedin.com/in/alexey-gorbachev-24b5305b/`

Nikolay Kulemin

Founder

Ph.D. in bioinformatics

University: Moscow Institute of Physics and Technology,
Department of Molecular and Biological Physics.

Specialist in bioinformatics and mathematical biology, has experience of academic research and applied developments in the genome industry. Nikolay is a founder of a company developing new algorithms for genomic analysis.

Email: nick@zenome.io
LinkedIn: `https://www.linkedin.com/in/nikolay-kulemin-50ab47a9/`

Vladimir Naumov

Genomic data scientist

Data scientist in human genomics

University: Pirogov Russian National Research Medical University

Bioinformatics and data analysis specialist. ex-CSO at iBinom, bioinformatics scientist at GERO, 8 years in genomics industry. Works on creating pipelines and developing new ways in analysis and visualisations of genomic data.

Email: vov@zenome.io
LinkedIn: `https://www.linkedin.com/in/vladimir-naumov-8285a25b/`

Sergey Popov

Developer

Blockchain, P2P, distributed systems, pure mathematics

University: Moscow Institute of Physics and Technology,
Department of General and Applied Physics

Has experience of theoretical physics, theoretical informatics, pure mathematics, development of distributed systems and smart-contract developing.

Email: sp@zenome.io

Dmitry Kwon

Advisor

Ph.D. in molecular biology, business development manager

University: Moscow State University, department of molecular biology

Specialist in molecular genetics. Dmitry has a significant scientific background, expertise genetic analysis technologies, genomic and diagnostic markets, has successful business experience in biotechnology and Dx top intl companies.

Email: dk@zenome.io
LinkedIn: `https://www.linkedin.com/in/dmitry-kwon-2763b119/`

Contents

Chapter 3 - Technical Part

Introduction: Genome Tokenomics

In the following a general description will be presented of the Zenome platform. The need for tokens and benefits for the prospective investor are discussed.

For the most part, genomic information is stored in databases, financed by governments or large corporations. Individually, each database contains insufficient data to make the quantum leap towards an era of genomics and precision medicine. At the same time, each database contains so much information that it's impossible for a single company to process all of that.

It appears that the exchange of genetic information is of crucial importance. The prospective genetic market must ensure protection from possible misuse and genetic discrimination in particular. It is particularly important to maintain transparency and equal access to this market.

Global exchange of genetic information should address the following issues:

- The fragmentation of genetic data.
- The limited access of scientists, medics and companies to genetic data.
- Low affordability of genetic testing.
- The lack of privacy of those who agreed to open access sharing of their genomic data.
- Insufficient computing resources.

Zenome aims at creating the personal genomics infrastructure, that would enable participants to:

- Upload genetic information and take control of it.

- Securely store own genetic information.
- Make profit by selling access to genetic data or part of it.
- Undergo genetic testing in exchange to the right to use genetic information.
- Get individual dietary recommendations or personal training program based on genetic makeup.
- Make use of other genetic services.

The principal customers of genetic information are companies interested in genetic targeting such as Google, Facebook, Unilever and pharmaceutical companies.

Inside Zenome Platform different types of information, namely genomic, personal and financial data, are inextricably intertwined. The specific nature of each type determines the way to store information of that kind. Financial data, which includes records of transactions, is stored on blockchain. Anonymized genomic data is stored on the distributed network. Participant's personal data is kept on their own computers only. Treating data of different kind diffently provides privacy as well as scalability of the system.

Since all data transactions, including buying and selling data, are governed by smart-contracts, reflecting the decentralized nature of the platform, interactions can only include balances stored on blockchains. Using any of pre-existing tokens for this purpose would result in unreasonable dependence on the valuation of external token or coin. Thus, a separate utility token should be issued to power economic interactions on the platform. This, in particular, means that you cannot buy a genetic data with «normal» money, you have to obtain tokens first.

Zenome DNA (ZNA) is an utility token on Zenome platform. The valuation of ZNA is tied to the success of the platform.

In this whitepaper we will discuss in further detail the most pressing problems in the area of genomics, as well as the solution Zenome platform involves.

Chapter 1

Genomics

1.1 Background

In this section, definitions of the terms “genome” and “genomics” are given. The history of the first genomic sequencing efforts and the emergence of high-throughput sequencing technology (NGS) are considered. The main manufacturers of reagents and equipment used to obtain genomic data are described. Issues regarding accumulating genomic data and reducing the cost of analyses are considered. A review of current genomic databases is provided.

Genome

The genome is the complete set of genetic instructions found in a cell [1].

The genome contains biological information necessary for the development and functioning of an organism. The human genome consists of linear double-helical DNA molecules organized into 22 pairs of chromosomes plus two sex chromosomes – X and Y. All information contained in a genome is encoded using quaternary code through a sequence of 4 nucleotides designated A, T, C, and G. The term ”to read a genome” means “to determine a nucleotide sequence by a sequencing process” [2].

The individual sequence of a genome defines a variety of organismal features, including appearance, susceptibility to certain diseases, athletic ability, metabolism, nutritional preferences, compatibility with sexual partners (the ability to conceive children), and many more.

The International Human Genome Project

The International Human Genome Project[1] was launched under the supervision of the NIH (National Institutes of Health) in 1990 to determine the complete sequence of the haploid human genome. The initial project leader was one of the discoverers of the structure of DNA, Nobel prize winner James Watson

A draft sequence of the human genome was completed in the middle of 2000 and published in the beginning of 2001 in the journal Nature. The cost of this international project completed with public funding was approximately $3 billion. In 1998, a private company, Celera Genomics, joined the race to sequence the human genome. The leader of the private project, which was developed in parallel with governmental institutions, was famous scientist and entrepreneur Craig Venter, who managed to raise $300 million in private investments for Celera's project. By using the new shotgun sequencing approach and more productive computational methods, the sequence of Craig Venter's genome was published almost simultaneously with the data produced by the international consortium in 2001[3] in the journal Science. The "full" human genome was published in 2007, and some human genomic regions that are difficult to sequence remain unknown.

Development of genome analysis

Extensive investments, a large number of outstanding participants from the scientific community, and competition among private and public organizations have provided considerable impetus for developing genome analysis technologies. As a result, modern sequencing technologies such as NGS (next-generation sequencing)[2] have emerged together with a new branch of science called bioinformatics, a young field of research at the intersection of mathematics, IT, and biology, that develops techniques and algorithms for the analysis of large biological datasets in productive and computationally effective ways.

The emergence of second and third generation sequencing technologies (NGS) has led to a strong reduction of genome analysis cost. While even in 2009 the cost of a full analysis of a genome was about 100,000 USD, currently the average price for the same analysis has dropped down to approximately less than 1,000 USD (see Fig. 1 and Table 3).

[1] https://en.wikipedia.org/wiki/Human_Genome_Project

[2] https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-

https://en.wikipedia.org/wiki/Moore's_law

Table 1: Equipment developers and reagent suppliers for genomic sequencing. Market capitalization taken from Yahoo Finance.

Company	**Products**	**Capital-ization**	**Country**
Illumina	Hardware, reagents, consumables, software	28.06 B	USA
Thermo Fisher Scientific	Hardware, reagents, consumables, software (a part of business)	68.98 B	USA
Oxford Nanopore Technologies	Hardware, reagents, consumables	534.41 M	UK
Pacific BioScience	Hardware, reagents, consumables	436,93 M	USA
Roche	Hardware, reagents, consumables, software (a part of business)	213.44 B	Switzerland
Agilent Technologies	Hardware, reagents, consumables, software (a part of business)	19.32 B	USA

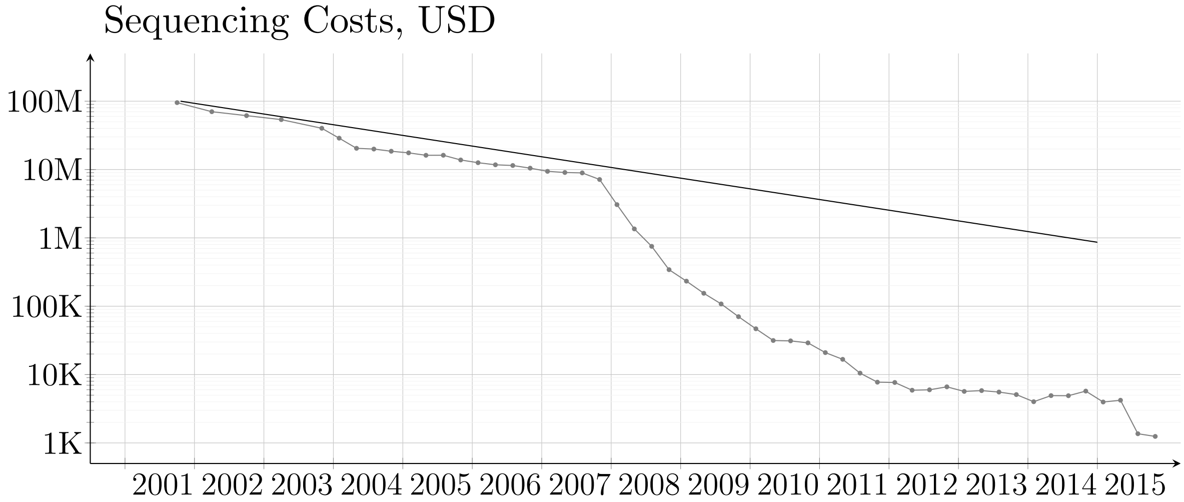


Figure 1: Moore's Law and cost reduction of genomic analysis. A significant drop in prices in 2008 was due to the advent of next generation sequencing technologies (NGS)[4].

Impact on other sciences

The development of genomics (the field of science studying various genomes) has led to the transformation of many scientific fields, from biology and anthropology to medicine and even social sciences. A number of top commercial companies such as Google, Apple, IBM, Amazon, and Alibaba have set the objective of using genomics to adjust their products and services according to the genomic profiles of their customers. Such adjustments will allow these companies to fine tune user relations and to predict their customers' needs and potential activities [3].

Genetic databases

The reduction in the cost of sequencing has led to an exponential increase in available genomic data. For example, a complete human genome in so-called ”raw data” format can represent 50 GB to 2 TB of data (depending on the sequencing depth required). To store such a large amount of genomic data, special genomic databases have been created containing various types of data, such as raw data obtained from genomic sequencers (”reads” or ”readings”), sequences of genes and proteins, sets of coding regions of a genome called exomes, and even the sequences of whole genomes (scaffolds); some of these databases contain clinically relevant information as well as the relationships between genetic traits and diseases. Most of these databases are centrally managed and financed by governments or large corporations. Scientists worldwide are involved in the addition of new data to these databases, enabling rapid updating and synchronization. In Table 2, some such well-known databases are described.

The majority of such databases are hosted in developed countries and are centrally managed and controlled by governments. Access to some of these databases is restricted, even for the scientific community, or is limited by commercial subscriptions. Although the founders of genomic databases claim they securely and anonymously store genomic data, in reality, the data stored are only pseudonymous, as in some cases, individuals have been identified based on their genomic information[5].

1.2 Genomic market overview

In this section, a brief analysis of the genomic technologies market is given. Examples of the most popular genomic products and companies providing various genomic services are described.

Genomic market growth dynamics

The market for genomic technologies is growing rapidly and is highly promising. Currently, the total market volume is approximately $25 billion with nearly tenfold growth, from $5.9 billion in 2010 to $60 billion in 2020 (predicted).

[3] `https://www.smeal.psu.edu/fcfe/documents/innovations-in-medical-genomics-pdf`

Table 2: Genomic databases

GenBank	http://exac.broadinstitute.org
Owner:	NCBI-NIH, USA
Product:	Genome sequences database
Stored:	More than 199,341,377 different genome sequences
ExaC	www.ncbi.nlm.nih.gov/genbank
Owner:	Broad Institute of MIT and Harvard, USA, ODC Open Database License (ODbL)
Product:	Exome Aggregation Consortium
Stored:	60,706 human exome samples/sequences
UniprotKB	www.ebi.ac.uk/uniprot/
Owner:	EMBL-EBI, SIB, PIR, UK, Switzerland, USA
Product:	Open Knowledge Base. Manual expert curation. Proteins and genes sequences.
Stored:	More than 555,100 manually reviewed and annotated record
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
Owner:	NCBI-NIH, USA
Product:	freely available archive for interpretations of clinical significance of genomic variants for reported condition
Stored:	>158 000 submitted interpretations, representing >125 000 variants.
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php
Owner:	QiaGen
Product:	Commercial database
Stored:	208,368 human mutation records with annotations
SNPedia	https://www.snpedia.com/index.php/SNPedia
Owner:	Open database
Product:	SNPedia is a wiki investigating human genetics
Stored:	107,073 SNPs and linked records
1000 Genomes Project	www.1000genomes.org
Owner:	EMBL-EBI, Wellcome Trust
Goal:	find most genetic variants with frequencies of at least 1%
Stored:	More than 2,504 Genome samples/sequences
100000 Genomes Project	http://www.genomicsengland.co.uk/
Owner:	NHS, Government of UK
Product:	UK government database containing a sequence of 100,000 genomes
Stored:	32,642 Whole genome sequences

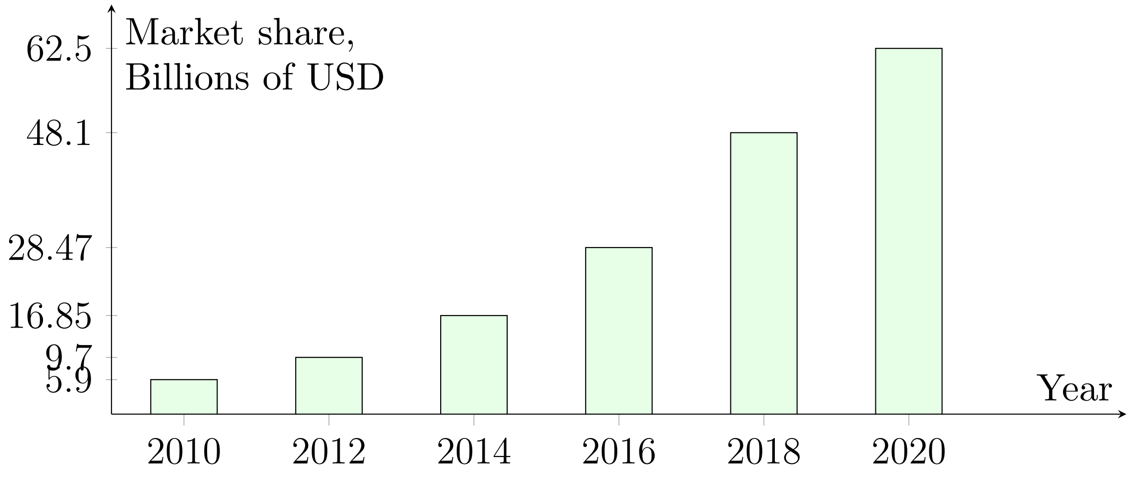


Figure 2: Genomic market growth dynamics for the period 2010-2024.

Product areas

As a key component of Industry 4.0, genomics has a broad range of potential applications in almost all economic fields. Main product areas of present genomics market:

- **NIPT (Non-Invasive Prenatal Testing)**

Description:	Based on DNA of fetal origin circulating in the maternal blood. Testing can potentially identify fetal aneuploidy[6] and gender of a fetus as early as six weeks into a pregnancy.
Market Share:	4 billions USD
Market Leaders:	Illumina, Natera, Ariosa, Sequenom.

- **PGS (Preimplantation genetic screening)**

Description:	Based on DNA-microarray or NGS-sequencing genetic profiling of embryos prior to implantation during IVF (in vitro fertilisation) procedure
Market Share:	336.4M USD
Market Leaders:	Illumina, Agilent Technologies

- **DTC (Direct-to-consumer) genetic testing, SNP-genotyping**

Description:	Genetic test based on DNA-microarray SNP-genotyping for getting some following recommendations: ancestry, nutrition needs, carrier status, optimal exercise.
Market Share:	2 Billions USD
Market Leaders:	AncestryDNA, 23andMe, DNAfit, deCode genetics.

- **Diagnostics company (Including oncogenomics)**

Description:	Different types of diagnostic procedures, genetics included. Sequencing of genes panels, Exomes, Whole-genomes, Liquid biopsy
Market Share:	16 Billions USD
Market Leaders:	Pathway Genomics, Human Longevity, Inc, Laboratory Corporation of America, Quest Diagnostics.

1.3 Challenges for Genomics

In this section, the main problems with contemporary genomics are considered. These problems must be solved to implement the concept of Genomics 2.0: the ubiquitous expansion and application of personal genomic technologies.

Low availability of genome analysis

The costs for different types of genome analysis are listed in the Table 3 and, in general, starts from $100. This price scale is extremely low relatively to the cost of genome reading 5 or 10 years ago[7].

However, the price of sequencing and bioinformatic interpretation remains high enough that providing the general public with access to genomic analysis is difficult, especially in developing countries.

Table 3: Genome analysis types.

DNA microarray	
Description:	analysis of 1-5 million pre-selected SNPs
Price range:	$100-500
Information amount	0.033% of full genome
Service providers	23andMe, AncestryDNA, DNAfit.
Genes panels	
Price range:	$100-2000
Information amount	0.001-1% of full genome
Service providers	Pathway Genomics, CeGaT.
Exome sequencing	
Description:	Sequencing for the coding part of a genome (exome)
Price range:	$250-3000
Information amount	2% of full genome
Service providers	BGI, CeGaT.
Whole Genome Sequencing	
Price range:	$600-10000
Information amount	80-98% of full genome
Service providers	BGI, FullGenomes, Human Longevity.

Note: actually, a part of a genome will not be sequenced. Its size depends on the sequencing depth and the details of sample preparation process. So the term "Whole Genome Sequencing" actually means obtaining the sequence of slightly more than 80% of a genome.

Obtaining a large amount of genomic information (sequences of genomes together with phenotypic characteristics) for residents of developing countries is extremely important from the perspective of obtaining the widest possible diversity of genomic information that, in turn, will significantly stimulate the development of the genomic big data market. Even in the mostly developed countries, less than 2% of the population has undergone any genomic analysis (microarray, exome sequencing, or whole-genome sequencing).

Privacy compromising as a price for participation in open biomedical investigations.

Informed consent for the collection and processing of personal data is a key issue in every biomedical research study. Any project involving human genome studies starts with the collection of signatures from individuals attesting to the fact that they understand the consequences and agree with the terms of the research study. The form of such consent varies depending on the project and may include giving permission to use the data in future projects, the consequences of which can be unpredictable.

In the Personal Genome Project, supervised by Harvard Medical School, participants voluntarily agree that their data and samples of their genetic material can be used multiple times and can be made available to other laboratories. Participants of this project are specifically informed that their identity could be de-anonymized and that their private data could become public. This project aims to ensure that genomic data from as many people as possible will be openly available to stimulate new research and development in the genomics industry. The authors of the project believe that if we do not provide open access to genomic data and information exchange, we are at risk of ending up with thousands of isolated, privately stored collections of genomic data (from pharmaceutical companies, genomic corporations, and scientific centers), but each of these separate databases will not contain sufficient data to enable breakthrough discoveries.

Conducting international multicentral scientific and clinical studies.

To perform new research and development in the field of genomics, it is necessary to conduct scientific and clinical studies on large samples from various population groups. Currently, collecting genomic samples from individuals with various ethnic backgrounds is difficult, as special projects need to be created, expeditions undertaken, and permission obtained from local regulators.

Today, only one startup is attempting to address this problem `https://www.dnasimple.org/` for a small fee and with the promise of anonymity.

Creation of biobanks storing materials from various individuals

Biobanks serve as mediators between the donors of biological materials (blood samples, bone marrow samples, etc.) and researchers by processing the materials obtained and storing them for future use. In general, biobanks are a key tool for the progress of personalized (precision) medicine and drug development. One of the most important functions of biobanks is the collection of donor material for future use, including blood, bone marrow, and even germ cells.

Currently, biobanks are actively being developed in many countries. However, there is the possibility that the most valuable samples will belong to the most financially secure biobanks, for example, biobanks belonging to large pharmaceutical companies, which will lead to unequal access to biobank material for different categories of researchers. Therefore, there is a need to create biobanks in as many countries and cities as possible up to the limit of individual bio-reservoirs for each individual.

Processing: genome analysis rate is limited by its bioinformatic processing. Adjustment of analytical software for widespread application.

Currently, the speed of obtaining genomic data using genome analysis technologies (sequencing) is high and outstrips the speed of processing these data. If we consider any large-scale scientific study of a large number of genomes, the experimental steps of obtaining genomic information account for no more than 20% of the study duration, while data processing comprises a much greater part of the project. Here, data processing means the steps taken from obtaining raw sequencing data to interpreting results and searching for various associations.

Another problem with modern software is that genomic analysis software was made by scientists and for scientists and thus requires adjustments to be widely applicable among physicians and for general consumers. As genomic analyzers the size of a USB stick already exist, the use of a personal sequencer in the same manner as a personal computer can be considered not science fiction but, rather, realistic in the near future.

Genomic data interpretation: mathematical models of disease developing risks

Various models and algorithms are used to rank the risks of developing diseases based on genetic data. The primary types of these models are based on the type of inheritance considered: monogenic, polygenic and multifactorial. Assessing the risks of developing multifactorial or complex diseases requires accounting for the influence and interference of many genes as well as environmental factors. For a more detailed description of the available methods for assessing the risks of complex diseases, see[8, 9].

To develop a new model for assessing the risk of developing a disease, it is necessary to involve a large number of scientists. It is also necessary to investigate many publications regarding the disease being analyzed, to identify its type of inheritance, to determine the polymorphisms and mutations contributing to the development of the disease, and to develop ”genomic algebra,” that is, a set of rules for risk estimation. When a model has been established and validated in silico, clinical studies must be conducted to assess its applicability. This approach is currently the most accurate, but it is expensive in terms of the amount of time and effort required.

Genomic data interpretation: machine learning application.

The use of machine learning algorithms to assess the risks of multifactorial diseases is being extensively investigated, but thus far, due to the lack of a sufficient number of training samples, existing mathematical models developed by biological scientists outperform machine learning approaches.

However, machine learning is already being used to predict certain complex characteristics of the human body. An example is the appearance of prediction in the work of Craig Venter and colleagues. The essence of their work involved analyzing the genomes and approximately 30,000 facial data points from several thousand volunteers. Based on the data obtained, training samples for machine learning algorithms were built and dependencies between genomic traits and individual appearance were determined. Because of this work, machines have learned to accurately restore a person’s appearance based on his or her genomic data[10, 11].

The results of this project enable the prediction of the appearance of a criminal or of an unborn child during the early stages of pregnancy. By obtaining a blood sample from a pregnant woman and extracting fetal DNA from the blood, the appearance of an unborn child on his or her 18th birthday can be accurately predicted.

To implement this project, Craig Venter recruited one of the best machine learning specialists from Google, Franz Och, a star computer scientist known as the chief architect of Google Translate[12].

Currently, machine learning is not widely used for diseases, as very large and correctly structured samples are needed for training. The creation of a comprehensive database of human genomes, as well as the availability of detailed questionnaires reflecting individuals’ health statuses, can spur the development of computational training in genomics and will result in highly predictive accuracy in determining the risk of disease development. At the same time, these data will be public and available to all users of the system, excluding the possibility of their monopolization. This availability is extremely important, as the concentration of large amounts of data in corporations’ databases will result in monopolies in the field of genomic machine learning.

Secure storing of very large data

Personal data security is very important; we all try to prevent the theft of credit card data, insurance information, banking account numbers, and medical information of any type. The theft of genomic information may seem unimportant for many people today. However, it can lead to very serious consequences that are difficult to predict, for example, the possibility that a portion of an individual's genomic sequence could be synthesized and planted at the site of a crime or a terrorist act.

Current solutions for this problem involve encrypted storage on a central server, such as[13], `https://www.pathway.com/`, `https://www.23andme.com/`, or `http://www.humanlongevity.com/`. This type of "closed" data storage is relatively safe, but it prohibits the possibility of sharing data and providing access to scientists from all over the world, which is a critically important condition for the development of modern genomic science.

Another problem associated with storage concerns the large size of a genome itself, and the exponential growth of the amount of genomic data available, as more and more individuals are subjected to genome sequencing. In one study [14], by 2025, the total volume of stored genomic data (given the size of a complete genome is 100 Gb) is predicted to reach 40 exabytes per year and genomic storage is predicted to become one of the largest consumers of storage and information processing capacities.

Table 4: Four domains of Big Data in 2025.

Astronomy	
Acquisition	25 zetta-bytes/year
Storage	1 EB/year
Analysis	In situ data reduction; Real-time processing; Massive volumes;
Distribution	Dedicated lines from antennae to server (600 TB/s)
Twitter	
Acquisition	0.5-15 billion tweets/year
Storage	1-17 PB/year
Analysis	Topic and sentiment mining; Metadata analysis
Distribution	Small units of distribution
YouTube	
Acquisition	500-900 million hours/year
Storage	1-2 EB/year
Analysis	Limited requirements
Distribution	Major component of modern user's bandwidth (10 MB/s)
Genomics	
Acquisition	1 zetta-bases/year
Storage	2-40 EB/year
Analysis	Heterogeneous data and analysis; Variant calling, $\sim$ 2 trillion central processing unit (CPU) hours; All-pairs genome alignments, $\sim$ 10,000 trillion CPU hours.
Distribution	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

The need for common database with continuously updated questionnaires.

At the same time, the lack of a public database implemented based on the concept of distributed storage and using open-source software could lead to the total domination of this market area by companies such as Google or Amazon due to their powerful servers[4]. If corporations or pharmaceutical companies become monopolists in the field of genomic information, we will observe the slow and expensive development of medicine with current treatment approaches persisting instead of a future in which a disease could be prevented even before its development or at the earliest stages of its manifestation.

[4] https://cloud.google.com/genomics/

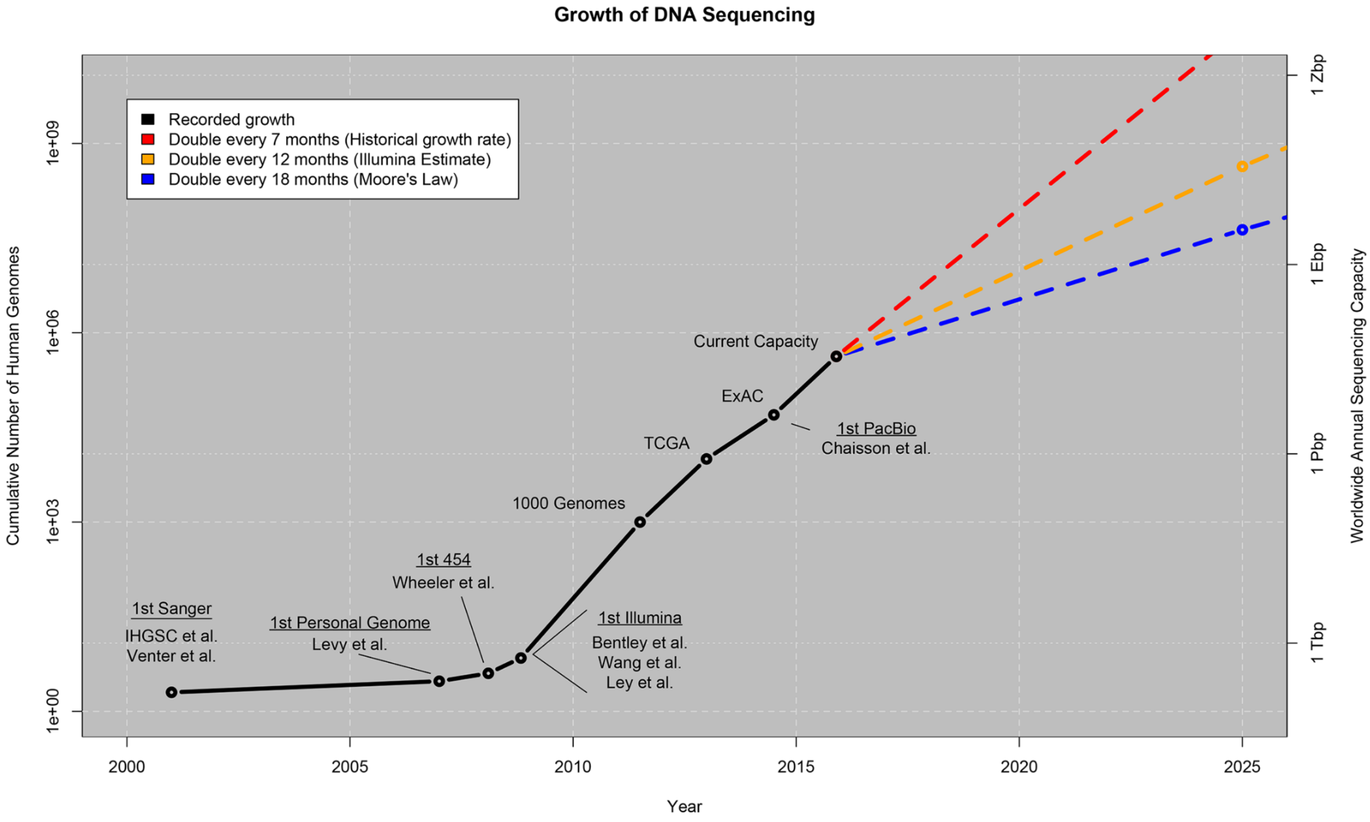


Figure 3: Growth of DNA sequencing data. This plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) and the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), and Zetta-basepairs (Zbp)). The values through 2015 are based on historical publication records with highlighted milestones in sequencing (first Sanger sequencing through the first PacBio human genome published) as well as three exemplary projects using large-scale sequencing: the 1000 Genomes Project, which aggregated hundreds of human genomes by 2012; The Cancer Genome Atlas (TCGA), which aggregated several thousand tumor/normal genome pairs; and the Exome Aggregation Consortium (ExAC), which aggregated over 60,000 human exomes. Many of the genomes sequenced to date represent whole exomes rather than whole genomes, but we expect this ratio to be increasingly biased toward whole genomes in the future. The values beyond 2015 represent our projections under three possible growth curves, as described in the main text. Taken from [14].

The near future of genomics

As a conclusion to this chapter, it is worth describing some hypothetical but technically feasible perspectives and dangers that could be faced by the genomic industry and society in general:

- A reduction in the cost of genome analysis and the miniaturization of genomic sequencing devices down to the level of cell phone plug-ins;
- The explosive growth of genomic data and its storage and the emergence of “genomic hackers” and privacy issues (protecting data);
- The blanket distribution of genomic medicine and telehealth;
- Changes to the food industry involving the implementation of personalized nutrition based on genomes;
- The development of personalized drug therapy;
- Dating based on genomic compatibility;
- Personality identification using genomic information, including the ability to make payments and obtain services;
- An increase in the average lifespan in all countries, the extension of active longevity, and a strong bias toward senior populations, late pregnancies, and decreasing birthrates;
- The development of genome-editing technologies;
- Appearance design for future children and other developments that are difficult to imagine. For example, the technical possibility of selecting healthy children and determining their future appearance both at the embryo stage and by obtaining fetal DNA from a pregnant woman’s blood already exists [10].

1.4 Ethical Issues of personal genomics

In this section, ethical problems associated with the development and ubiquitous distribution of genomic technologies are considered, such as privacy, public databases, open access for researchers, possible misuses and threats to personal liberties due to the expansion of genomics.

NOTE In the present white paper, we provide only a brief review of the main problems and perspectives of the genomic industry. For a more detailed discussion of privacy and data security problems, see https://www.smeal.psu.edu/fcfe/documents/innovations-in-medical-genomics-pdf.

Privacy

Personal genomic information is very sensitive for many people. However, many people do not fully understand that, based on their genomic information, it is possible to determine their lifespan, propensity to make emotional decisions (manipulability of decision-making), likelihood of developing various mental diseases and risk of sudden death due to, for example, heart arrhythmia.

Such information could be disadvantageous for job recruitment, election participation, and medical insurance pricing. There is also the possibility that a bad actor who knows the sequence of a genome could leave fragments of DNA identical to that genome at, for example, the location of a terrorist act to frame or illegally accuse someone. One could be denied medical treatment (or required to pay a higher fee) or barred from obtaining a desired job.

Corporations and governments could deliberately influence one's decisions and purchases using their knowledge of "weaknesses" in one's genomic information. Thus, the protection of genomic data privacy is necessary to protect the equal rights of various categories of people.

At the same time, some studies have been performed that allow the identification of individuals' identities based on their anonymous genomes [5, 15].

Moreover, some companies (http://www.humanlongevity.com/media/) possess machine learning-based algorithms that can accurately reconstruct the appearance of an individual using only his or her genomic data[16].

Publicity

Such an approach will enable the development of preventive medicine through which the analysis of large amounts of data allows the prediction of disease development (before its occurrence), enabling actions that increase lifespan and improve quality of life[17] as well as the identification of donors worldwide to meet various medical needs. Unfortunately, no valid solutions currently exist for the public use of genomic information while maintain individual privacy. However, some startups working in this field using blockchain technology should be mentioned.

Encrypgen, a startup that has recently conducted an ICO, described the existing problems and the relationship between privacy and publicity using blockchains[18]. However, the white paper associated with this project lacks a description of a technical implementation that would solve the problem of privacy and availability.

Another project in this field is the DNAbits startup[19], whose founder, Dror Samuel Brama, has patented a general approach to data storage and transfer using blockchain technologies[20]. However, this company has not technically implemented its concept during the past three years.

The right to own genomic data?

Currently, there is no legislative definition of the right to possess one's own genetic information. In some developed countries, including the USA, Germany, and Austria, citizens do not have the right to access and possess their genetic data in the context of its interpretation[21]. An agent, represented by a physician or medical center with the right to provide such information, is needed. This path is used by the companies Pathway Genomics in the USA and CeGaT in Germany (`http://www.cegat.de/en/`).

To undertake genetic analysis, the advice of a physician who could be a provider of genetic testing is needed, and only this physician has a right to interpret the information provided by genetic analysis.

In the USA, there are service providers in the field of "genetics for fun," such as the companies 23andMe and Ancesty.com, which sell genetic tests directly to the end customer, but these companies can only provide information on ethnic origin and certain health-related characteristics (for example, sports characteristics) and lack the permission to provide most medically valuable information. These restrictions, imposed by regulators such as the FDA, do not hinder the ability of 23andMe to sell access to genetic data to large pharmaceutical companies. Some such deals are known: a deal was made with Genentech (a subdivision of the pharmaceutical giant Roche) for $60 million[22] for a study of Parkinson's disease, and a deal was made with another large pharmaceutical company, Pfizer, for a study of inflammatory bowel diseases (e.g., Crohn's disease)[23]. Some reports also claim that 23andMe has had negotiations with Novartis over an Alzheimer's disease study[24].

Thus, we currently give large companies the right to manage our genomic information, to store it, and to profit from it. Corporations, behind the veil of good intentions, monopolize genomic big data, and we cannot predict how this monopoly will influence future drug prices and medicinal discoveries.

The right to access genomic information

Another ethical issue should be discussed; above, we noted that privacy violations and access to genomic information could be used illegally, for example, by an employer. One could be fired or denied a job promotion based on genetic information. For those with jobs related to the safety of people and systems, such as truck or bus drivers, pilots, atomic power station operators, or people with other similar occupations, health status is critically important, and genomic information could prevent an accident or even a disaster. In some professions, a potential danger could threaten the worker rather than bystanders, such as a coal or diamond miner with lung problems. For these cases, a discussion involving professionals and experts as well as the general public is necessary to develop legal standards regulating the use of genomic information by employers.

Chapter 2

Concept

2.1 Zenome Project

Philosophical View

Public awareness of genomic medicine remains quite low in the developed and even worse in developing countries. It means that people, in general, have little understanding of possible benefits of genomics as well as possible dangers associated with it. In many countries it resulted in developing overly-complex institutional procedures to protect genetic infromation from possible misuse that, on the other hand, hinders scientific progress.

The Zenome Platform is going to raise awareness on genomic medicine, so users can make conscious decisions regarding their data. To ensure that, the Zenome Platform is based on the following fundamental principles:

Individual ownership of personal genomic information Each participant has all rights for personal genomic data.

Freedom of choice Each participant decides how individual genetic information should be used. One may decide whether or not to participate in scientific/clinical research.

The right to share The participant may grant access to genetic information to a third party in a way restricting data copying.

Privacy Private data encryption makes it impossible to access individual genetic information without explicit user's permission.

Distributed data storage Distributed database architecture provides high availability and fault tolerance through replication and scale out ability.

Distributed data processing Data is processed on many network nodes at the same time. Any user can become a node by providing disk space and CPU time to the network.

Scalability The platform architecture enables great scalability and flexibility of the system.

The Zenome Platform Ecosystem

On the Zenome platform user[1] is engaged in many types of different interactions throughout the system. These interactions take place at different system levels, don't interfere with each other and involve different patterns of interactions. Thus, they should be represented as distinct entities that have different roles.

The following roles are available on the Zenome platform:

(Calculating / Storing) Node that provides storage and CPU power for a reward.

Person who has uploaded individual genetic data to the platform and possibly uses genetic services.

Analyst who is interested in analyzing genetic information on the platform. May represent: a data scientist, a scientific organization and so on.

Service Provider that implements a user-space genetic service (possibly paid) on the platform. Basically, it's an organization that uses genetic data as a part of its business.

IN SHORT

> **Each user is engaged in a number of interaction of different kind, taking on different roles. Some of them, such as Service Provider and Analyst , require special knowledges, but Node and Person don't.**

Every role will be discussed in details later.

The System Architecture's Overview

The Zenome platform is a distributed application that consists of 3 major layers.

[1] User in a broader sense, i.e. a person or software that runs on behalf of that person.

Network (and Data Access) Layer : provides a level of abstraction that encapsulates network interaction and provides interface to a distributed environment to upper layers.

	Blockchain	DHT Kademlia
Data Storage Cost	High	Low
Data Immutability	True	False
Performance	Low	High
Deterministic Result	True	False

This layer consists of two distributed systems of a completely different nature:

Distributed Ledger (based on blockchain) that records transactions between participants in a verifiable and permanent way. To access blockchain node's software runs embedded Ethereum client.

Distributed Hash Table Network (based on Kademlia protocol) that combines physical nodes into an overlay network and enables message passing between nodes and distributed data storage.

ROLE (Calculating / Storing) **Node** operates at the Network layer.

Middleware contains account management, consistent interface to security features of the underlying layer and high-level APIs for software that runs on Application level.

ROLE **Service Provider** operates upon Middleware. So-called External Services Platform API enables third-parties to start genetic services on the platform.

Application Layer Zenome application features an advanced end-user interface that translates user actions to Middleware in very consistent way. Interface is extandable by design so that genetic services can run natively on the software-stack.

Genetic services market

The market of genetic services is currently developing in the following areas:

1. **Research and technology adoption** into the market.
2. Providing **genomic diagnostic services**.
3. **Government certification** of genetic technologies.
4. **Developing a legal framework**. In particular, legislative measures to safeguard genetic information;

> NOTE
>
> The structure of the market is quite complex. Some players are, in fact, developing in several directions to find their place in the market in the face of rapidly increasing consumer needs.

The following players are represented on the genetic services market:

Scientific corporations are working on the discovery and adoption of new technologies on the market. These include:

- Pharmaceutical corporations, biotechnological and diagnostic companies, such as Pfizer and Myriad
- Companies that develop and sell all necessary chemical supplementary reagents (such as Life Technologies).

IT-bioinformatic companies are engaged in invention and development of the methods of computational data processing. Players in this sector are still struggling to deal with the types and volumes of the data obtained.

Scientific and medical centers are playing a leading role in the provision and development of genetic diagnostics services.

Commercial laboratories are providing a fast, efficient and usually relatively cheap genetic diagnostic services. They are possessing large financial and resource abilities.

Direct-to-Consumer genetic diagnostic companies increase the population's interest in genetic diagnostics. Currently this segment is very small, but in future, it can grow into one of the leading market part and can be adopted into clinical practice.

Table 5: Comparison with Similar Products on the Market

	We	GeneCoin	Encrypgen	23andMe	Pathway Genomics	Snpedia (Promethease)	Human longevity
Decentralized	✓	✓	✓	–	–	–	–
Suitable for non-human organisms	✓	✓	✓	–	–	–	–
Customer is the owner of his data	✓	✓	✓	–	–	✓	–
Possibility to load your own data	✓	✓	✓	–	–	✓	–
Opened nonprivate data	✓	✓	–	–	–	–	✓
Performs its own data analysis	✓	✓	–	✓	✓	✓	✓
Provides a report for customers	✓	–	–	✓	✓	✓	–
Uses AI and Machine Learning	✓	–	–	–	–	–	✓
Sharing without transmitting huge data	✓	–	✓	–	–	–	–
Earn using your data	✓	–	–	–	–	–	–
Opened for scientists	✓	–	–	–	–	✓	–
Is a platform for other tools	✓	–	✓	–	–	–	–

2.2 Roles in The Zenome Platform

From Resource Nodes' Perspectives

The system from a perspective of a computational node

Node — a participant providing the resources of his or her computer (storage and CPU time) for the purpose of distributed storage and processing of genetic data for a reward in ZNA tokens.

To become a computational node, user runs Zenome software on his or her computer and activates `Node` role using graphical user interface. The software must run continuously in the background.

NOTE A command line version of the Zenome software is also provided that can be used specifically to start computational Node.

The policy of system resources' allocation and task management can be flexibly configured by the node owner:

- maximal storage size that software is allowed to use
- a policy of computational resource usage:

  Fixed CPU/GPU usage number of cores and maximal loading (in percent) per each core.

  Dynamic CPU/GPU usage resources are allocated to the software in such a way that they do not interfere with user applications.

- Choose computational process (by id) that will have the highest execution priority.
- Shut down the node temporarily: request the network to transfer the data not available elsewhere to other nodes. Wait until data transfer is completed and disconnect.
- Shut down the node completely: request the network to transfer all data to other nodes, wait until data transfer is completed and delete data.

Person/User

Person — an individual willing to provide his or her genetic information to the Zenome system in order to make profit from selling his or her personal data or to use genetics-based services available on the platform.

A user installs Zenome software to work with personal genetic information.

Using the graphical user interface, user is able to:

- Create personal account
- Manage tokens: recieve, transfer, use, pay for data storage, spend on paid genetic services.
- Upload genetic data (file format is mostly detected automatically).
- Manage personal data: provide personal data, fill in a questionnaire, view a list of the most popular questionnaires.
- Work with targeted offers using safe suggestions subsystem.
- Use genetic services and configure privacy level for each service. individually.

When loading genetics data, the level of privacy can be chosen among:

Full Privacy In this case data are stored in encrypted form, and full price is charged for storing such data.

Standard Privacy Genetic data are stored as fragments making it impossible to identify the user. Each fragment is stored in the system overtly. Information regarding matching fragments to user IDs is private. In this case, the storage is cheaper due to subsidies.

The Public Access Data is stored overtly. Storage is free due to subsidies.

NOTE

Attention! Although there are no technical restrictions against increasing the level of privacy, only future data is affected. The data once made public would never be private again. Consider this when you upload the data first time.

Data Consumer

Data Consumer — a scientist, commercial company, scientific organization or other platform participant, who is interested in genetic data analysis using platform capacities. Data consumer is able to make requests to the users and set a refund that will be paid to the responding users

NOTE

Note: there are some restrictions for queries that are designed to prevent system user de-anonymization and other platform misuse. These restrictions also depend on consumer ranking in the system.

Service Provider

Service Provider — an organization that uses genetic data in its business and thereby implements a user service on the platform.

NOTE

User can choose himself or herself which data he wants to share with the service provider. User will be notified if the requested data could be used to identify him.

NOTE

For example, service provider cannot see more than 70% of mutation list in clear form using direct query or to get the information matching raw data (like fastq files) with user questionnaire

2.3 Genomic data

Types of omics (genomic) data

Within the framework of the concept, 3 types of omics (genomic) data may be distinguished depending on the payment and the information value:

- Open data, which is not valuable for its owners, but is important for scientists.
  Example: Genome of Helicobacter pylori bacteria strain.
- Open data, which is valuable both for its owner and for the consortium.
  Example: Genomes of the majority of network participants.
- Restricted data that are simply stored within the network.
  For restricted projects of various commercial and public institutions.

Preprocessing of genetic data

The handling of NGS genomic data (and most other omics datasets) usually consists of two independent steps:

1. **Preliminary processing of ”raw” data.**
2. **Dedicated analysis of genetic sequences** for the developing of personal recommendations or within the framework of a research.

NOTE

A reference genome — is a digital DNA dataset assembled as a representative example of a species’ set of genes.

Preprocessing of NGS genomic data includes the following steps:

1. Alignment of NGS reads to the reference genome.
2. Searching for mutations and other differences from the reference genome, and saving their list in gVCF format.

> NOTE
>
> *Note: The protocol is the same for the data of non-human organisms. Of course, in this case an appropriate reference genome is required.*

If personal dataset represents the result of microarray genotyping technology (23andMe file type), it can also be uploaded to the platform as gVCF since the data formats of these file types are similar.

The Fake Data Problem

If the genetic data of other (non-human) organism are uploaded instead of the correct genome (accidentally or deliberately), this will be detected during the preprocessing of raw data and the user will be notified. If the user intentionally uploads distorted (fake) genetic data into the system, this can be detected using many well-known verification methods before saving them into storage.

> NOTE
>
> The economic inducement to refrain from uploading fake data is that the payment for data storage should be made upfront for the whole year.

The problem of user identification based on genomic data

Open access to genetic information raises the problem of identifying users by their genomic and other data. If a user decides not to fully disclose his or her genetic information, appropriate measures should be taken at each step of processing and storing the data. To solve the problem of user identification, the interaction should be designed in such a way that at each stage no node could determine the ownership of the genetic material by specific individual, or even the city in which this individual lives in.

> NOTE
>
> The differences between residents of one city constitute approximately 0.01% of the sequence.

At each stage, the goal above is achieved in different ways:

1. In the preprocessing phase — by dividing the source file into parts so that the average coverage is lower than the confidence threshold (6 copies).
2. In the storage stage — storage account fragmented by length.

Storage of genomic data

Genomic data are located in a distributed network based on DHT Kademlia protocol. The participants who provide resources for the operation of this network (see details regarding `Node` role) receive payments for it in ZNA tokens. In order to receive payment, they need to prove to the network that they are really storing these data. The procedure of this checking is based on blockchain usage. Encryption is used when necessary.

NOTE *Note: Integration with Storj and FileCoin nodes will also be implemented.*

As already mentioned, the data could be raw and processed.

Type of data	Raw	Processed
	Features	
Format	fastq / bam	(lists of mutations) gvcf / vcf + bed / 23me(txt)
Size, Gb	50	2
Value	To improve the technology of sequencing and processing (for the equipment development market)	To conduct research, as well as to make a report.
	Storage conditions	
Number of copies	At least 3 in the independent nodes	At least 5 in the independent nodes

Genomic data are stored divided into fragments in such a way that the length of a fragment does not allow to unambiguously identify that this fragment belongs to a specific individual.

NOTE *Information regarding which genome fragments constitute the user genome is also private and can be obtained only with the user's permission.*

2.4 Personal profiles

Filling personal questionnaires significantly increases the applicability of genomic data. The users fill in the questionnaire using the graphical interface.

NOTE

If some questionnaire becomes popular, the application prompts the user to fill it. Each analyst may create his or her own questionnaire and place it into the platform.

Specification of the questionnaire

A number of questionnaires can be huge; therefore, it is necessary to introduce the concept of the questionnaire specification.

Specification of the questionnaire — is a complete description of all fields of the questionnaire and the allowed values of these fields.

NOTE

Formally, the specification of the questionnaire contains a reference to the author, a description and an ordered set of records, each of which corresponds to one field of the questionnaire.

Fields may have several types:

Numeric field The value of the field is a integer.

Multiple choice The value of the field is a number of an answer.

NOTE

The answers to these types of questions will be put in open access (without any reference to the user), since they do not threaten the privacy.

String field The value of the field is a string. It is a private field, because it potentially allows compromising the user identity based on a specific answer.

Private block Allows making any set of fields private, regardless of their real type.

Queries to the system.

Statistical data regarding certain genomic statuses will be open to public if the owner has not decided to encrypt them. In addition, the information regarding the available answers to the questionnaires is open too. Therefore, everyone knows, for example, the number of network users 25 years of age or having a mutation rs6025 (coagulation factor V).

The architecture of the system makes it impossible to extract the full database:

- During creation of associative inquiries, the customer does not get access to raw data.
- Basic fee includes only a limited number of requests per day. The fee for additional queries included in the quota grows exponentially during a day.
- If the result of the associative query contains less than 100 users, the result will not be provided to the customer.

If some user completely encrypts his or her data, then he or she decides on his or her own to whom the personal data, including the fragments of which his or her genome consist of, are allowed to be transferred. No analyst will be able to know what types of data are encrypted.

Secure transfer of personal data

The process of transferring personal data between participants in the system should possess the following properties:

1. Full data to be transmitted should be available only to the buyer and to the seller.
2. Transmission of tokens should only take place if the data have been transmitted successfully.
3. An attempt to sell incorrect data should be identified and blocked.
4. An attempt to deliberately falsely accuse the seller in selling of incorrect data should be revealed.
5. Data transmission should not be trusted to some third party.

> NOTE The blockchain technology will be used for secure data transfer. However, it should be considered that storing (and transferring) large amounts of information to the blockchain is resource consuming. Therefore:

6. It is allowed to transfer only a small amount of data through the blockchain. The remaining data can be transmitted through a simple encrypted communication channel.

2.5 Rating system

The platform will create rating for:

- Separately for each genetic fragment or block of personal data
- Organizations
- Service providers
- Data providers (such as DNA sequencing labs)

NOTE

> There will be no individual user rating, since it actually represents the sum of his or her genetic fragment ratings, which is not available to public. If the genetic data was uploaded using Organization account, then the initial rating will be automatically increased by the organization's rating.

The factors affecting the rating of the genetic fragments

Confirmation from the laboratory increases the rating of uploaded data in proportion to the rating of the lab in which they have been obtained. Confirmation may represent a digital signature of the laboratory, or data uploading by the laboratory itself upon client's request.

Check for plausibility Allows to check user genetic information using predefined statistical models of polymorphism frequencies and genetic linkage. This module is under development.

Participation in research The rating increases with the number of successful researches involving the fragment. If the result of the information investigation have been considered implausible, the rating decreases.

2.6 Use cases

Individual user

For the individual users there is an opportunity to obtain their genomic information and to turn it into a source of income. The combination of the genome and its interaction with the environment is a valuable information resource. Our platform will allow the user to safely manage this resource.

The platform provides ability to securely store and share genetic information, allowing users to receive a variety of genetic services. Here are a few examples:

- reports and recommendations on nutrition, risks of diseases, cosmetology, diet, fitness
- search for relatives and ancestry clarification
- dating services
- individual selection of clothing, shoes, home climate setup, travel destinations and areas of residence
- different variations of the genetic reports for a group of individuals, for example sports teams or working groups
- almost every aspect of human live is influenced by genetics, so let us see what new companies can come up with using our platform

Besides the plurality of services, a user is given an opportunity to make profit from his or her genetic uniqueness by providing questionnaire data to the companies for research purposes. Thus, individual data together with genetic information become an analogue to commodities or mineral resources

Health

Modern healthcare and personalized medicine cannot be imagined without the use of genomic technologies. The platform will allow patients to securely share genetic information used in the clinic with medical personnel:

- the individual dosage and intolerance to the drugs (for example, individual dosing of the anticoagulant warfarin based on the genetic characteristics)
- personal acceptable ranges of biochemical body parameters (for example, PSA marker)
- genetic predisposition to various diseases (for example, a high risk of macular degeneration and the need for additional research and prevention)
- Transplantation and organ donation. Users may securely share the information regarding the type of their HLA antigens that determines the compatibility between individuals during transplantation. Thus, it will be possible to create a secure database of donors and volunteers to save lives through transplantation.

Company

There are two types of Companies, first of which provides users with services based on genomic data, while the second is interested in obtaining genetic data from users to conduct their own research.

First type is described in user's use cases zone. Second type can be described as buying users genomic data to conduct their own research and to improve the consumer properties of products, genetic targeting of products and advertising, some examples of which are:

- For example, a pharmaceutical company is planning to release a new drug that acts against the mutant cancer protein. Company may find system users, who survived the disease, to pay them for genetic data and to get the frequency of mutations in a gene encoding the protein that is targeted by active substance
- consumer company plans to enter a new market and needs to test how users perceive the taste of the product. It is known that some flavoring causes resentment among carriers of a particular genetic variant of the gustatory receptor gene. The company sends over a network an offer to study the carriers of this genetic variant and picks up another flavoring, or finds out the frequency of this genetic variant in different markets and thus targets the product for different markets.

The scientific community

For the scientific community the system opens the possibility of storing, sharing and performing research with various genomic data. Because the platform is not limited to working with human genomes, it can be used for the secure storage and processing of genomic data, for example, relevant to agriculture (plants, animals, microorganisms).

In general, the presence of ecosystems leads to the enrichment of the scientific community through access to general population data, even without reference to the individual questionnaires. In addition, with the consent of the users, they can also become a part of scientific research.

The platform also provides distributed computational power, access to which will allow to process large volumes of genetic data (similar to AWS adapted to work with genetic data).

Chapter 3

Technical Part

3.1 Distributed objects

A concept of distributed subsystem

(Distributed) Subsystem — a collection of some fraction of the system (platform) elements and processes that can be represented from the object-oriented perspective as an entity that has a distinct identity and exhibits a well-defined externally visible behavior.

To give a detailed characteristic of a subsystem, its following aspects should be described:

1. **Structure**: the elements and processes forming the system.
2. **External Behavior**: interaction of a subsystem as a whole with other participants. In particular:
    - **Interfaces**: a collection of possible queries to a subsystem as a whole.
    - **Actions**: actions taken by a subsystem concerning the other system participants.
3. **Internal State**: the aggregate internal state of a subsystem.

Subsystem may be represented as **quasiobject**, with which other participants are able to interact.

NOTE

> The prefix «quasi-» means that real interaction takes place with subsystem elements, which, in turn, have complex interactions with each other as a part of this real interaction, so that all these stuff could be considered as interaction

with some aggregate object.

NOTE *Below we will drop distinctions between a subsystem and its representation as a quasiobject.*

Interactions with a subsystem can be represented as a sequence of relatively small number of basic operations:

- Subsystem interface, that is, the actions of other participants concerning subsystem.
- The actions concerning other platform participants.
- Internal processes that change the internal state of a system.

Internal processes

Internal processes in a broad sense — represent a collection of all internal processes of each subsystem element and interactions between these elements.

Internal processes (in a narrow sense) — represent processes within subsystem as a whole which change its internal state. Full description of internal processes contains all subsystem behavior excluding the issues of its specific implementation.

3.2 Main distributed subsystems of the platform

The platform has the following organization levels:

- Basic System Layer (Critical Infrastructure)
- Data Storage and Processing Level
- High-Level Interactions

The following subsystems are included within the platform:

1. Basic System Layer

   Low-Level Interoperation Basic subsystem of message exchange between network nodes. It also enables creation of distributed hash tables.

Authorizations The infrastructure of accounts and access management to private information.

2. Data Storage and Processing Level

   Storages Abstraction level for access to distributed file system.

   Processing Infrastructure for a distributed calculations.

3. High-Level Interactions

   Secure queries A subsystem that enables formation of the offers for buying genomic data which could be shown to appropriate users only.

   Open data operations Provides tools for operation with open (not private) data.

   External Services Platform Provides API for connecting external centrally managed services to the platform. Arranges secure data transfer using conventional web protocols.

3.3 Low-Level Interoperation Layer

Distributed P2P network

A basis of the platform is P2P network based on Kademlia protocol. The network involves the computers of users who have installed Zenome software.

Node — is a node of distributed P2P network that represents a user's computer with Zenome software installed on it.

NOTE

> Consequently, an overlay network is constructed between the devices participating in the network. It represents a virtual network in which to each participant a 'NodeId' is assigned, that has no relationship with real IP-address of the device. Each node stores the list of "proximal" nodes, where the distance between nodes is calculated based on nodes' 'NodeId' and is not related to common distance.

The nodes store the data by using distributed hash tables.

Distributed network implementation characteristics

A modified protocol specification is used. Main differences are the following:

- Nodes can exchange arbitrary messages between each other. A node is able to transfer a message to other node knowing its ‘NodeId‘ only.
- Several hash tables may exist within the network. Hash table is identified by string key.
- Different policies of value storing and deleting can be set for different tables.
- The data transferred from one node to another are encrypted (see below).

Message exchange in a distributed environment

Thus, participants of peer-to-peer network are able to exchange the following messages:

`PING`	Verify that a node is still alive
`STORE(T,K,V)`	Store the value ‘V‘ by the key ‘K‘ in the table ‘T‘ in a node receiving the message
`FIND_NODE(N)`	A node receiving the message will sent the data regarding the nodes which are closest to the node ‘N‘ among the nodes it knows.
`FIND_VALUE(T,K)`	If the pair ‘(K,V)‘ is stored in a receiving node, send the value ‘V‘, else send the data regarding known nodes which are “closer” to file.
`SEND(M,N,D?)`	Sends the message ‘M‘ that can also contain the data ‘D‘ to the node ‘N‘. ‘FIND_NODE‘ is used to locate the node.

NOTE

> Note: this interaction level is a transport level of the platform.

3.4 Blockchain

Basic information

The platform uses Ethereum blockchain that represents a single decentralized virtual machine (EVM). The desired system logic can be implemented using smart contracts.

QUOTE

> A contract is a collection of code (its functions) and data (its state) that resides at a specific address on the Ethereum blockchain.
>
> — Introduction to Smart Contracts (Solidity manual)

Thus, smart contracts are able to store data. For the data in the blockchain the abstract type system described above is still valid (but with some restrictions).

A subsystem for working with the Blockchain

> NOTE
>
> Attention: implementation details could be different depending on the platform used. The description below is relevant for PC platform.

To provide an access to the Blockchain, node software includes the full implementation of Ethereum Node providing access by means of ‘JSON-RPC 2.0‘.

Secret keys are stored in the encrypted storage. A user is prompted to set password at first application launch.

> NOTE
>
> Although it is technically feasible to use an existing account, it is recommended to create new one.*

The application interface allows to:

1. Create new account
2. Import an existing account
3. Set up the backup of private storage

> NOTE
>
> Attention: It is recommended to set up the cloud backup since this will allow to keep an access to account even in the case when the computer will be physically unavailable.

> NOTE
>
> *Software allows getting access to Ethereum node command line. This function is primarily intended for debugging. It is not recommended to use a command line if you do not understand its purposes.*

Internal tokens

Economical interactions within the system are provided using internal ZNA tokens. They represent valid Ethereum tokens and can be bought and sold at exchange houses.

Account concept

Account in the platform allows user to interact with the system by playing several roles concurrently. Each role in account corresponds to separate smart contract in the blockchain.

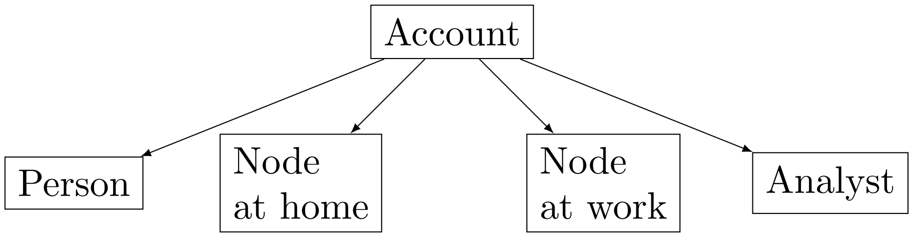


Figure 4: The account of a user who has provided his or her genome (Person role) to the platform, who works as a bioinformatician and supports two computational nodes located at home and at work.

Role — is a characteristic of the system participant regarding the specific type of interactions in which he or she takes part. Each participant may play several roles simultaneously.

NOTE A separate subsystem is responsible for the role distribution.

3.5 Distributed data storage network

The principles of distributed network operation

The distributed network based on DHT Kademlia protocol is used for storing data of arbitrary types. The nodes of this network are represented by registered resource providers, which earn a reward for storing the data and performing the computations needed. The participant who has uploaded the data pays a reward for data storing, but, in a number of practically important cases these charges are partly or fully subsidized by the system. For example, there exists a mechanism of subsidizing the storing of scientifically valuable data.

The storage unit is represented by an arbitrary data block and unique key by which it can be accessed.

NOTE The size of a data block is restricted to prevent misuses.

External services support. A possibility of integration with Storj and FileCoin nodes, as well as with centrally managed storages, will also be implemented.

Data storage reliability assurance

To assure the reliability of data storage, in view of the fact that the nodes can freely connect to and disconnect from the network, the data are stored independently on several nodes simultaneously. The type of data stored determines the number of nodes on which the information duplicates.

Table 6: Exemplary parameters of genetic data.

	Raw	Processed
Format	fastq/bam	gvcf/vcf + bed/23me(txt)
Size	50Gb	2Gb
Nodes	≥ 3	≥ 5

When the number of the nodes changes, the data stored are redistributed to satisfy the requirement for the minimal number of the nodes.

Since everyone who wants to can become a node in the distributed network, it is not safe to believe that the node really stores the file when it claims so. Thus, the check of data retrievability is periodically performed according to the corresponding protocol from Security Layer. The data on the checking results is introduced to the blockchain and can be the reason for issuing a reward or changing node rating.

Data storage privacy assurance

The privacy of data storage in such a distributed network becomes possible due to application of asymmetrical cryptography. The actual method of encryption is determined by the corresponding security protocol.

NOTE

> It should be noted that subsidizing mechanism could not be applied for storing the encrypted data.

Specific features of storing genomic information

The belonging of some genome fragment to the particular individual can be unambiguously identified if the fragment is rather large. This is why the genomes in the distributed network are stored as rather short fragments.

Genome fragmentation is always performed based on the reference genome. For each reference genome (for example, another version of human reference genome or a genome of the organism of different species) a fragmentation is chosen only once, and each fragment is assigned with identifier that is unique for this reference genome.

3.6 User's personal data

Record — is a minimal unit of the data exchange. It is not possible to transfer (for example, to sell for a reward) a part of the information from a record. It is possible to create a new record that will contain only a part of the data, but in such a case this new record will be significantly less interesting for potential buyers due to its low rating.

The data scheme determines which information and in which format should record contain. Scheme usage is a flexible tool for the unification of data exchange format between network participants. Data scheme identifier could be any string that enables individual to find its description in the network, for example, URL or smart contract address with a description. It should be understood that primary purpose of the data scheme is to unify the conception of what is included in the data between the buyers and the sellers of these data.

Questionnaire specification

Filling personal questionnaires increases the value of genomic data substantially. A huge number of questionnaires may exist, so it is necessary to introduce the concept of questionnaire specification.

Questionnaire specification — represents a complete description of all fields of the questionnaire and the allowed values of these fields.

Formally, a questionnaire specification contains the author reference, a description and ordered list of records each of which corresponds to one field of the questionnaire. **Numeric field** A user's answer should be within the allowed value range $[a, b]$. The answer is coded by the unsigned integer, where the count starts from left border of the range:

> NOTE
>
> An answer to such type of question is publicly available since it does not threaten the privacy.

Multiple choice The user's answer is coded as unsigned integer representing serial number of the answer in a list. If the value equals zero, then the user has preferred not no answer to this question.

NOTE

Attention: an answer to such type of question is publicly available since it does not threaten the privacy.

String answer The user's answer is stored as a string.

NOTE

An answer to such type of question is a private information.

Filled questionnaire — is a data structure containing the user's answers to the questions of the questionnaire.

NOTE

This structure is fully encrypted if necessary. It is stored at least on the user's computer. A user can upload the encrypted backup to the blockchain if he wants to.

Personal data request from users have some distinctive features:

- Not all users satisfy the criteria of the specific research. To check whether the data are appropriate, an access to personal information is required.
- Personal information should not be transferred outside the user's computer neither explicitly (directly) or implicitly ("warrant canary").
- In a case when user satisfies the criteria, he or she will receive an offer to share his or her personal data for a reward.

CONCL. 1

Since the information transfer is not permitted, the checking should be performed in isolated environment. The code executable in this environment have a full access to private information but cannot interact with other parts of the system.

CONCL. 2

The result of this isolate function's execution is not permitted to be sent as an answer to potential customer since it also threatens the privacy.

CONCL. 3 It is reasonable to choose the restricted set of personal data, which are necessary to make a conclusion on a separate checking step. The user will see on what data the decision is based on the exchange offer screen.

CONCL. 4 The data to be transferred explicitly are listed on the exchange offer screen. Only the data that have been requested to make a checking could be transferred implicitly, and the list of such data will also be available to the user.

Personal data request:

- **In the first step** an access is requested to personal user data.

  Only the data identifiers of which were listed in this request will be available to checking function.

- In the second step, the code of the checking function is executed within an isolated environment, and its execution result represents the formulated offer regarding data exchange or offer cancellation. In the latter case no data are transferred anywhere.

  In the former case a user is notified when the exchange offer is formulated, then he or she examines the offer, the list of data used during checking and the list of data that will be transferred if the user agrees to the offer.

  If the user rejects the offer, no data are transferred.

  If the user accepts the offer, only the data explicitly shown to the user are transferred.

Bibliography

[1] NHGRI. *Talking Glossary of Genetic Terms. Word «Genome».*
URL: `https://www.genome.gov/glossary/index.cfm?id=90`.

[2] Venter J.C., Smith H.O., Adams M.D. "The Sequence of the Human Genome". In: *Clinical Chemistry* 61.9 (2015), pp. 1207–1208.
URL: `http://clinchem.aaccjnls.org/content/61/9/1207.long`.

[3] Adams M.D. Venter J.C. Smith H.O. "The Sequence of the Human Genome". In: *Science* 291.5507 (2001), pp. 1304–1351. ISSN: 0036-8075.
DOI: `10.1126/science.1058040`.
URL: `http://science.sciencemag.org/content/291/5507/1304`.

[4] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).*
URL: `www.genome.gov/sequencingcostsdata`.

[5] Melissa Gymrek et al. "Identifying Personal Genomes by Surname Inference". In: *Science* 339.6117 (2013), pp. 321–324. ISSN: 0036-8075.
DOI: `10.1126/science.1229566`.
URL: `http://science.sciencemag.org/content/339/6117/321`.

[6] H. Christina Fan et al. "Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood". In: *Proceedings of the National Academy of Sciences* 105.42 (2008), pp. 16266–16271.
DOI: `10.1073/pnas.0808319105`.
URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2562413/`.

[7] Sboner, Andrea and Mu, Xinmeng Jasmine and Greenbaum, Dov and Auerbach, Raymond K. and Gerstein, Mark B. "The real cost of sequencing: higher than you think!" In: *Genome Biology* 12.8 (Aug. 2011), p. 125. ISSN: 1474-760X.
DOI: `10.1186/gb-2011-12-8-125`.
URL: `https://doi.org/10.1186/gb-2011-12-8-125`.

[8] Rachel R. J. Kalf et al. "Variations in predicted risks in personal genome testing for common complex diseases". In: *Genet Med* 16.1 (Jan. 2014), pp. 85–91. ISSN: 1098-3600.
DOI: `10.1038/gim.2013.80`.
URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3883880/`.

[9] Karen Norrgard. "Calculation of Complex Disease Risk". In: *Nature* (2008). URL: https://www.nature.com/scitable/topicpage/calculation-of-complex-disease-risk-756.

[10] Kathryn Nave. "How Craig Venter is fighting ageing with genome sequencing". In: *WIRED UK* (2016). URL: http://www.wired.co.uk/article/craig-venter-human-longevity-genome-diseases-ageing.

[11] Amalio Telenti et al. "Deep sequencing of 10,000 human genomes". In: *Proceedings of the National Academy of Sciences* 113.42 (2016), pp. 11901–11906. DOI: 10.1073/pnas.1613365113. eprint: http://www.pnas.org/content/113/42/11901.full.pdf. URL: http://www.pnas.org/content/113/42/11901.abstract.

[12] Luke Timmerman. "Google Translate Star Leaves Venter's Human Longevity For Illumina-Backed Grail". In: *Forbes* (2016). URL: https://www.forbes.com/sites/luketimmerman/2016/09/27/google-translate-star-leaves-venters-human-longevity-for-illumina-backed-grail.

[13] João Sá Sousa et al. "Efficient and secure outsourcing of genomic data storage". In: *BMC Medical Genomics* 10.2 (July 2017), p. 46. ISSN: 1755-8794. DOI: 10.1186/s12920-017-0275-0. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547444/.

[14] Zachary D. Stephens et al. "Big Data: Astronomical or Genomical?" In: *PLOS Biology* 13.7 (July 2015), pp. 1–11. DOI: 10.1371/journal.pbio.1002195. URL: https://doi.org/10.1371/journal.pbio.1002195.

[15] Erika Check Hayden. "Privacy protections: The genome hacker". In: *Nature* (2013). URL: http://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940.

[16] Laure-Anne Pessina. "Reconstructing a face from DNA: an EPFL alumnus takes the stage at the 2016 TED Conference". In: *School of Engineering (Federal Institute of Technology Lausanne)* (2016). URL: http://sti.epfl.ch/page-129921-en.html.

[17] Melanie Swan. "Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen". In: *Journal of Personalized Medicine* 2.3 (2012), pp. 93–118. ISSN: 2075-4426. DOI: 10.3390/jpm2030093. URL: http://www.mdpi.com/2075-4426/2/3/93.

[18] Patrick Lin. “Blockchain: The Missing Link Between Genomics and Privacy?” In: *Forbes* (2017).
URL: https://www.forbes.com/sites/patricklin/2017/05/08/blockchain-the-missing-link-between-genomics-and-privacy.

[19] Justin Zimmerman. “DNA Block Chain Project Boosts Research, Preserves Patient Anonymity”. In: *CoinDesk* (2014).
URL: https://www.coindesk.com/israels-dna-bits-moves-beyond-currency-with-genes-blockchain/.

[20] D.S. Brama. “Method, System and Program Product for Transferring Genetic and Health Data”. US Patent App. 14/218,865. July 2015.
URL: https://www.google.com/patents/US20150205929.

[21] Melanie Swan. *Blockchain: Blueprint for a New Economy.*
URL: https://www.goodreads.com/book/show/24714901-blockchain.

[22] Matthew Herper. “Surprise! With $60 Million Genentech Deal, 23andMe Has A Business Plan”. In: *Forbes* (2015).
URL: https://www.forbes.com/sites/matthewherper/2015/01/06/surprise-with-60-million-genentech-deal-23andme-has-a-business-plan.

[23] “23andMe, Pfizer to Launch Inflammatory Bowel Disease Genetics Study”. In: *GenomeWeb* (2014).
URL: https://www.genomeweb.com/clinical-genomics/23andme-pfizer-launch-inflammatory-bowel-disease-genetics-study.

[24] “Genetic Wild West: 23andMe Raw Data Contains 75 Alzheimer’s Mutations”. In: *Alzforum* (2017).
URL: http://www.alzforum.org/news/community-news/genetic-wild-west-23andme-raw-data-contains-75-alzheimers-mutations.