

# The Hutch Report

[www.thehutchreport.com](http://www.thehutchreport.com)

# The Genomics Bottleneck

## Who holds the solution?

Insights, Ideas and Tools for the New Economy

### GENOMICS AND BIG DATA

A look at the flood of data to come from Genomics along with the challenges it will bring

### MPEG-G THE COMING STANDARD

Our in depth and exclusive look at those attempting to provide a standard to the benefit of the whole Genomics Industry

### GENOMICS AND THE BLOCKCHAIN

We look at 12 companies that are applying blockchain technology to Genomics

### SWOT ANALYSIS

A balanced view of the industry through our analysis of its strengths, weaknesses, opportunities and threats

# The Hutch Report

The Hutch Report was founded in 2015. Originally from North America, the founders have lived in Europe for the past 20 years. We are headquartered in Geneva, Switzerland a global hub for diplomacy, banking and biotech.

Together we have worked with a number of technology startups and gained a wealth of experience, one having been sold to PayPal (Nasdaq: PYPL) and another to Macrovision (Now TIVO, Nasdaq: TIVO). In that time we have compiled an extensive network of contacts in the European and North American startup communities as well as the venture capital and private banking sectors.

Through research, interviews and analysis we are out to discover next great opportunities, investments and nuggets of knowledge. All of the content posted on our website, in the blog and in our feature reports, is unique and original. The Hutch Report is frequently engaged by a variety of organizations including hedge funds for industry insights as well as directly by a broad range of companies for marketing and management consultancy. If you are interested to know more, please feel free to use the contact form or drop us an email as noted on the contact page. Any other feedback, comments and suggestions are also welcome.

## **Special Update:**

In order to capitalize on the findings and opportunities we have uncovered, we are currently evaluating the initiation of a seed investment fund of \$30M to \$50M in order to help accelerate the growth of a select few early stage Genomics related startups. We ask accredited investors to contact us at [Genomics@thehutchreport.com](mailto:Genomics@thehutchreport.com) with "Seed Fund" in the subject heading.

For more content and exclusive downloads, visit  
[www.thehutchreport.com](http://www.thehutchreport.com)

# "INCREASING THE PACE OF SCIENTIFIC DISCOVERY"

Completing the sequencing of the first human genome in 2003 was a key breakthrough that took more than 10 years and cost over US \$3 billion. Since then, the speed at which genomes can be sequenced has more than doubled, easily outpacing Moore's Law. Today's high-throughput sequencing machines can process the human genome in a matter of hours at a cost approaching US \$1,000. These advancements have allowed researchers to run more genome analysis in less time, greatly increasing the pace of scientific discovery.

## In this Report

### Genomics and the Big Data Challenge

You have no doubt received mail in the last few years letting you know that your personal data—credit cards, bank accounts, health, insurance—may have been accessed by someone who violated the security of the organization you trusted to keep your information safe? Those letters could become even more alarming if they're reporting the theft of your genomic data. These are a few of the challenges we address.

### MPEG-G The coming standard

One thousand petabytes of genomic data are already being stored worldwide, and the aggregate cost of genomic storage is expected to grow from US \$0.5 billion today to US \$5 billion by 2021. There will be a need to reduce the footprint of genomic datasets in FASTQ and BAM, at the same time preserving genotyping accuracy in order to help reduce hardware storage costs.

### Genomics and the Blockchain

Genomic data privacy and security will be an absolute must in the future. Blockchains have strong privacy advantages for genomic data. A blockchain is essentially a public record that, in this case, would track every time your data is used. That record is hack-proof because it is in a chronologically encrypted chain that is authenticated and validated by many users and can't be altered by others. But will it be the solution the industry is looking for?



## editor's note

Welcome to our 6th feature edition of The Hutch Report! Our last report on Quantum Computing received some tremendous feedback, along with a number of requests from readers interested in the intersection of quantum computing and genomics.

This seems like a potentially powerful convergence given that quantum computing techniques enable the ability to solve equations dealing with mass amounts of data and that genomics is expected to generate mass amounts of data that will need to be processed. We decided that a Hutch Report on genomics would be merited. Therefore, we are pleased to present you with our look at the world of genomics and the associated big data issues.

As a reminder we leave the academics to the academics, The Hutch Report work is approached with the markets in mind. Our goal is to provide those interested in the markets and business an additional and objective lens through which to evaluate risks and opportunities in particular areas.

*The Hutch Report  
Team*

# INTRODUCTION TO GENOMICS

## A COMPLEX FIELD MADE UP OF MULTI-DISCIPLINARY TECHNOLOGIES

Genomics is a vast field of science involving many different disciplines with the focus on fully understanding how an organism's DNA works (its structure, function(s) and how they are mapped) and can eventually be modified. A complete set of an organism's DNA, including all of its genes, is referred to as a genome. Genomics may sometimes be confused with genetics but these are different.

Genetics looks at individual genes and how these are passed down through generations whereas genomics is looking at the entire DNA structure and complete set of genes. Both fields are of course tightly linked and it is perhaps sometimes a disservice by naming or trying to categorize scientific (or artistic) endeavours and restricting them to certain areas whereas the actual science and research itself can be quite vast and exceed boundaries which may be implied by a name.

The study of genomics is receiving more and more attention from investors as well as the general public. There is a good reason for this. Genomics impacts the very core of our lives and all living things around us. The human genome, that is, its entire DNA sequence, contains up to a total of 20,000 to 25,000 genes.

There are about 6,000 human diseases caused by faulty genes. In a recent video segment on Real Vision Television, Cathie Wood estimates that today we are able to solve about 5% of monogenic (single gene) diseases. With gene editing that could go up perhaps to 100% for monogenic conditions.

According to market analysts that represents about a two trillion-dollar opportunity. The market opportunity for solving polygenic conditions is conceivably higher. In addition to humans, genomic engineering can be deployed in agriculture and farming to enable faster and healthier production of food for the growing population.



# Deoxyribonucleic acid

To situate and better understand genomics it helps to at least have a basic overview of DNA. DNA is short for deoxyribonucleic acid. So it is a chemical substance, a type of acid. It is a chemical substance, a type of acid and is located in the nucleus of every cell in a living organism. There are also smaller amounts of DNA contained in some of the outer parts of a cell in what is called the mitochondria which is part of the cell structure that generates energy for the cell. DNA contains instructions for the replication of cells and these are also passed from adult organisms to offspring during reproduction.

Through a complex process, enzymes are able to “read” the information in a DNA molecule, which through interactions with other molecules and acids, notably RNA, gets translated into amino acids and proteins. Proteins are just another type of molecule, but very important ones. Proteins provide many crucial functions including providing structure to every cell, regulating body processes such as breaking down foods, providing energy, creation of hormones, transportation of other molecules (for example hemoglobin, which transports oxygen throughout the body).



A DNA fingerprint sequence that displays the DNA genetic chromosome makeup of an individual person.

Proteins also form antibodies that help prevent infection, illness and disease by identifying and assisting in destroying bacteria and viruses. So, any defect in DNA may result in a suite of many ensuing complications if the resulting protein is not correctly created or doing its job correctly. In addition to DNA, defective RNA is also implicated in a number of human diseases including heart disease, some cancers and strokes. As we understand more about DNA, researchers are also understanding more about the role(s) that RNA plays as well.

In addition to the treatment of diseases and conditions, as with any new advances in science and technology there are moral and legal questions which society will have to answer. It will become even more delicate when the technology advances to a state where features such as intelligence, eye colour, hair colour, muscular structure and strength, and height for example can be “chosen”.



The technology is not there yet and is probably still a long time away, however, it is not difficult to imagine the day when it will be possible to choose your child's eye colour. We have barely solved all the moral, legal and financial issues with existing health care and health insurance so it would be naïve to think all of the moral, legal and financial issues will be solved by the time genomic health care is available. Aside from moral questions there are also questions and concerns on privacy. Proposals have been made for creation of a universal DNA database which could be used by governments in a variety of ways that include diverse uses such as to better plan health care investments for their citizens or even crime fighting, much like a universal fingerprint database. Opponents of state run DNA databases are deeply concerned about the potential misuses and security breaches by malicious factions. Every one of us has slightly different DNA except in some cases (identical twins at least when they are born). Everyone gets a random mix of DNA, approximately 50% from each parent and this mix includes also a random mix of mutations. Even though that sounds like such a mix would create a large percentage of difference in the DNA of each human, remarkably we all share basically about 99.9% of the same DNA as everyone else. And just as remarkable, while that 0.1% difference sounds low, given the vast quantity of DNA that 0.1% difference actually is quite a lot and makes each of us distinctly unique. Genomics is a complex field and as scientists and researchers develop tools for this science there are many multi-disciplinary technologies that are required.

This edition of the Hutch Report seeks to discuss and provide a view into the major technology themes that are rapidly evolving as the field of genomics advances. We will be producing additional short segment papers on particular companies for some of these themes. In these segments we will attempt to identify which companies may stand the best chances of becoming the winners.

# Major Themes



**DNA Sequencing**

*(also referred to as Genome sequencing)*

**Data compression**

**DNA editing**

**Chemogenomics**

**Robotics**

**Computational Genomics**

**Machine learning in genetics and genomics**

**Big data and cloud computing in genomics**

**Biological Nanotechnology**

*(Biological nanorobots, DNA- based nanothermometers, bio computers, etc)*

**Legal and compliance**



# THE GENOMICS VALUE CHAIN

Genomics is a discipline which analyses the function and structure of genomes (the complete set of DNA within a single cell of an organism). The genomics value chain describes the process by which genomic samples are transformed into useable information to guide and develop treatments or improve patient care. It can be divided into the following five stages:

**Sampling** - The process of extracting, cleansing and transporting DNA (e.g., blood or saliva samples). Overall it is considered a low-value area since it does not necessarily require clinicians to complete it. Although this is where all the DNA data is derived from.

**Sequencing** - The process of decoding the order of the nucleotides in a genome is called sequencing. The sequencing process has been made more efficient in the last few years by the development and use of high-tech machinery. The ability to sequence the genome on a large scale is the reason for the rapidly decreasing costs.

**Analysis** - This stage enables us to understand whether the sequence of nucleotides reveals any variation when compared to other genomes. Once DNA has been sequenced it can hold a variety of data forms. By performing analysis using software and other methods, this information can be standardised, compared, and areas for investigation can be identified.

**Interpretation** - Interpretation is the process of translating analysed genomic information into insights for clinicians and pharmaceutical companies. Clinicians should be able to make treatment decisions based on this interpretation. It is currently the smallest of the sub-segments.

**Application** - Genomic information is used to provide diagnostic treatments, targeted therapies or drug development. The main users of applied genomics are pharmaceutical companies and, in the long-term, healthcare systems and clinicians. This section will require significant data volumes and sufficient skilled workers to develop to the attainable level.

|                | SUB-SECTORS       | CAPABILITIES   | DESCRIPTION   | VALUE  |
|----------------|-------------------|--|---|--|
| SAMPLING       | N/A               | Clinician-dependent, although some companies allow consumers to send samples by post | The process of collecting and packaging samples (e.g. saliva, blood). The kits used to collect DNA samples are fairly simple. | Low: can be performed with basic medical equipment. Complex supply chain as samples need to be stored and shipped appropriately specially. |
| SEQUENCING     | EXTRACTION        | Manufacturing  | Decoding the order of the nucleotides in a genome. DNA sequencing on a large scale is done by high-tech machines.             | High but limited headroom: backbone of genomic analysis but hardware may become commoditised   |
|                | CONSUMABLES       |  |   |  |
|                | INSTRUMENTS       |  |   |  |
| ANALYSIS       | DATA CLEANSING    | Software   | The process to identify disease-causing variants, often run by bioinformatics software.                                       | Significant value in locally developing software and creating databases to continually refine this information                             |
|                | VARIANT CALLING   |  |   |  |
|                | DATA SERVICE      |  |   |  |
| INTERPRETATION | REPORTING         | Understanding of healthcare, relatively manual                                       | Taking analysed information and providing clinically useful interpretations and results                                       | High: this is an added-value service that directly caters to the needs of key healthcare system and pharmaceutical buyers                  |
|                | LINK WITH EHRs    |  |   |  |
|                | TAILORING RESULTS |  |   |  |
| APPLICATION    | DRUG DEVELOPMENT  | Pharmaceutical or clinical expertise   | The process of directly using genomic information to improve targeting of clinical services                                   | Significant value in targeted sectors e.g. personalised medicine and advancing oncogenomics  |
|                | CLINICAL SERVICES |  |   |  |
|                | DIAGNOSTICS       |  |   |  |

# The Genomics Bottleneck

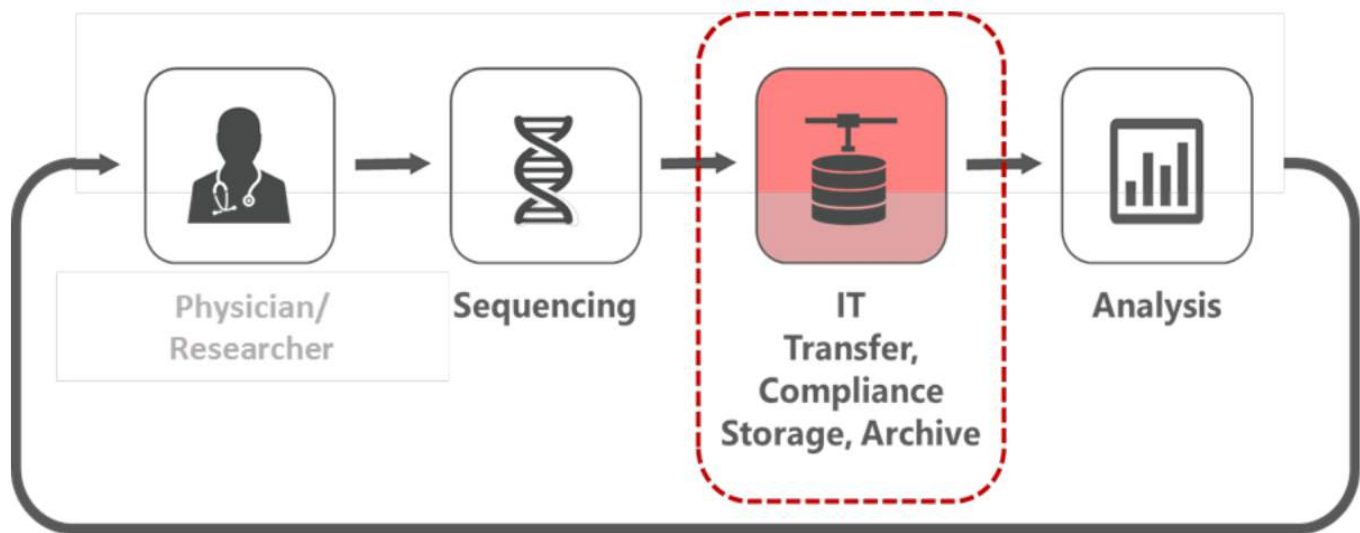


The life sciences and healthcare industries are in the midst of a dramatic transformation that will make personalized medicine common-place. Every segment of the sequencing value chain, from DNA sequencing itself, to analysis and diagnostics, to the support services that operate across different segments, is attracting interest and investment.

Thanks to the improvement in sequencing technologies, there has been a drastic reduction in costs. This has encouraged increased sequencing activity which has in turn caused a dramatic data explosion in genomics.

Genomic data is in fact, one of the fastest growing datasets in the world. Each genome can be composed of billions of nucleotides stored as plain text files. A recent Intel analysis stated that it would take 7.3 zettabytes, meaning  $7.3 \times 10^{21}$  bytes of data to store the genomes of our global population. This is equal to 50 percent of all data on the internet in 2016 and does not factor in the data created when analysing and using this information.

We have seen the computational power speed of computers and the internet increase greatly over the years. Regardless, there is still a lack of computational infrastructure. The genomics industry will still require a more powerful, secure, scalable, and sustainable technical infrastructure if it is to meet the unprecedented demands for data storage and computation that will be needed to support future initiatives.



Genome scientists currently rely on solutions that cannot possibly support future large-scale genomics initiatives. For example, current data sharing methods are inefficient and compromise security. Data is shared between organizations over file-based transfer protocols (e.g., ftp, sftp, fasp), requiring large volumes of data to be downloaded over the internet onto a local hard drive before it can be used, which promotes data redundancy and consumes expensive network bandwidth and data storage capacity.

The repeated transferring of data out of a single secured environment reduces the ability for gatekeepers of sensitive, identifiable genomics information to oversee access to researchers or audit their use of it. The centralization of data and infrastructure in a small number of institutions restricts access by the wider community. Current data processing workflows are developed in-house and hardcoded to run on specific local hardware architectures. This limits their reproducibility and confounds the comparison of results.

---

INSIDE THE WORLD OF

# Genomics and Big Data



We were quite struck that the quest into this nanoscopic world is generating massive amounts of data. As mentioned earlier, a driving reason is the reduction in cost of DNA sequencing. It is estimated that the sequencing of the very first human genome had a price tag of about US \$300M in 1999 to 2000. In 2006 the cost of sequencing a human genome came in at around US \$14M. In 2015 the cost was down to around US \$4,000 and today in 2018 the cost for generating a high quality human genome sequence is US\$1,000 and under.

With this decrease in cost it is also yielding continuously increasing massive amount of data, big data, which is already starting to pose challenges in research, scientific and commercial efforts.

Some of these challenges will likely be solved by an increase in capacity in underlying computer and hardware infrastructure. Other challenges may be solved by software solutions and standardization.

In discussing the challenges of big data, it is useful to consider the lifecycle of the data itself. A common way of phrasing this lifecycle is as follows:

- Data Generation and Storage
- Data Sharing
- Data Processing and Analysis

Big data from genomics, and other fields of study (Astronomy, Astrophysics, Meteorology, etc) as well as commercial businesses (think Netflix, Google, Facebook, Amazon, etc) is pushing the envelope on current technological capacities for data storage, data networking and sharing, and formatting for analytics and processing. In the following section, a few observations are presented in an attempt to develop a viewpoint as to where the interesting opportunities are developing around the big data challenge for genomics and through association, potentially with any other endeavours or fields of study relying on big data.

# Data Generation and Storage

How much data is being generated by genomics? Starting from the fundamentals, human genome sequencing is the process of determining a complete DNA sequence of an organism's complete set of genetic instructions within a cell. These instructions are made of four nucleotides which are the building blocks for the nucleic acids in DNA. These four nucleotides are Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). Students of computer science or binary math will recall that four values can be represented by two binary bits. In this case of the four nucleotides, each could be represented as two bits: A=00, T=01, G=10, C=11. A base pair represents the binding of two of these nucleotides together. Therefore, one base pair can be represented with 4 bits. There are 3.2 billion base pairs in one human genome. The result is that one complete human genome could be represented with 12.8 billion bits. In practice, only 6.4 billion bits are needed because researchers have observed that the nucleotides always bind with each other very specifically. An A always binds with a T and a G always binds with a C. Because of this binding of specific pairs, only one base (two bits) in the base pair needs to be counted, instead of two bases (four bits). The result is that only 6.4 billion bits are needed to store the 3.2 billion base pairs. Dividing 6.4 billion by eight we can derive the number of bytes, then dividing that result by 1024 we get the number of kilobytes and if we divide again by 1024 we get the number of megabytes required to store the human genome. The result is that approximately 800 megabytes are required to store one human genome. 800 megabytes for a human genome in itself does not seem like a very big amount of data, certainly not these days.



So why all the fuss? This is because this is a theoretical calculation and assumes that an entire genome can be perfectly sequenced as a single chain of 3.2 billion symbols. In reality modern sequencing devices are able to produce several hundred million “reads” (fragments of the whole genome) per day. According to the specific technology employed, each read can contain from a few dozens to several thousand bases together with optional metadata. Clinical applications require that in order to be reasonably sure that the reconstructed genome is actually correct, a redundancy in the order of 30 to 50 is necessary in the reads produced by sequencing devices. Therefore, the data that needs to be stored at this point is already much larger than 800 MB. Genomic data needs to be aligned and then compared with other data including reference genomes, along with a mass of other data including physical conditions of the organisms being studied.

When aligning and comparing 3.2 billion base pairs with other data, the required amount of processing is significant. Researchers and scientists are examining a diverse array of areas including gene regulation, genome evolution, microbe colonies in humans, genetic differences of cancer within different patients and how it impacts patients differently and the list goes on. Due to the promise of genetic medicine it is estimated that as many as 2 billion people may have their genomes sequenced by 2025.

In addition to human cells, there is an estimated ratio of 3:1 of bacterial cells to human cells within each of us. These bacterial cells comprise the microbial flora within humans and are intricately linked to human health and disease. Estimates are that there are over 120 trillion microbial cells in a human adult. These organisms also have their own DNA which would need to be understood in order to fully understand human health.

The Human Microbiome Project (HMP) was an NIH initiative which ran from 2008 and closed in 2017 and worked on, among other things, sequencing a portion of these bacterial cells. The HMP resulted in hosting 2.3 TB of compressed genomic data.

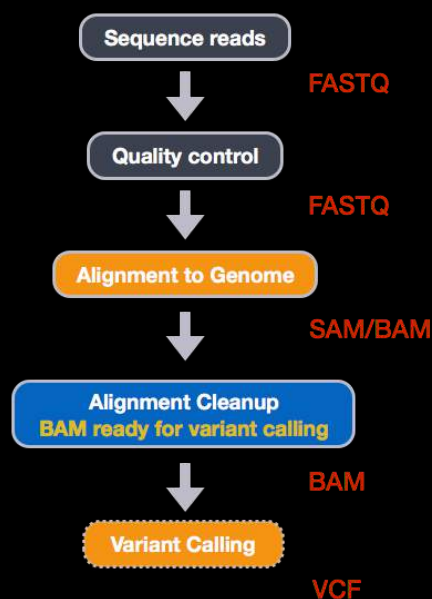
Beyond the human genome, there are over 1.2 million other species of plants and animals on the planet. Sequencing of other species will be used for research into agricultural (think food and farming), energy, and environmental reasons. For example, in China, one organization has already sequenced over 3,000 varieties of rice and are working on sequencing additional plant and animal genomes. Estimates have shown that since 2004 the amount of genomic data being generated is doubling every year. No one has a complete grasp of the entire amount of genomic data already generated and stored to date because of the highly distributed nature of the creation of the data in over 55 countries and thousands of labs each storing its own data.

The largest genomic data repository in the EU, the European Bioinformatics Institute (EBI) had over two petabytes of data in 2013. With the amount doubling each year that would imply they now store over 64 petabytes of genomic data as of the writing of this article. One petabyte is estimated to be the equivalent of 20 million 4-door filing cabinets full of text or 500 billion pages of standard printed text.



As mentioned above, during the sequencing of genomes, there is much more data that gets generated and needs to be stored. In order to be useful, the original data, the genetic soup, must be read several times. As with any process which involves going from an analogic or physical to digital format there is a lot of noise, corrupt data, which must be filtered out so the more reads that are made increases the ability to filter out noise and errors as well as identify possible mutations. After the genome is read it must be aligned to a reference genome in order to be able to map it and identify relevant variations. After it is aligned it is then compared for variants. A file containing an aligned whole human genome sequence ends up weighing approximately 330 GB. At this point there are various formats for storing and subsequently processing and working with the data as well as differences in how much of a genome is actually read or stored.

The typical flow of raw sequencing reads, with some of the current and common data formats used at each step until the variant identification stage, is depicted here:



#### Units of data

- Bit (a 0 or a 1)
- Byte (eight Bits)
- Kilobyte (1024 Bytes)
- Megabyte (1024 Kilobytes)
- Gigabyte (1,024 Megabytes, or 1,048,576 Kilobytes)
- Terabyte (1,024 Gigabytes)
- Petabyte (1,024 Terabytes, or 1,048,576 Gigabytes)
- Exabyte (1,024 Petabytes)
- Zettabyte (1,024 Exabytes)
- Yottabyte (1,204 Zettabytes, or 1,208,925,819,614,629,174,706,176)

With some back of the envelope calculations, the current cost for cloud storage including accessing the data (Amazon, Microsoft, Google, Rackspace, etc) is approximately US \$0.021 per gigabyte per month for large data volumes of relatively frequently accessed data, otherwise referred to as “hot” storage. Costs can be less if data is in cold storage and not needed to be accessed frequently or access bandwidth is not needed. One petabyte of data stored on the cloud would cost about US \$22,000 a month. It is estimated that by 2025, genomics will require anywhere from 2 to 40 Exabytes of storage per year. At current market rates for cloud storage fees the market for genomics data storage would range from US \$1 billion (cold) to US \$20 billion (hot) in storage fees per year. According to RightScale, from information they published in 2017, the costs on average would be about US \$25/year per whole human genome sequence.

Genomics is of course not the only field dealing with the storage and generation of big data.

Other examples of where big data are being generated and requires storage includes scientific endeavours such as astronomy, meteorology, astrophysics as well as commercial businesses such as YouTube, Netflix, Twitter, Facebook, and the list goes on. We could not find precise statistics, but various articles on the web estimate the amount of data storage used by Netflix for their entire catalog is about 2.75 petabytes of compressed data.

In comparison to large external hard drives developed by Seagate and Western Digital, the Samsung 2.5 inch is the world's largest capacity solid state drive with 30 TB storage capacity. With that capacity it would require about 35'000 such hard drives to store one Exabyte of data. At the time of this writing the price of the Samsung 30TB hard drive is not known, however, if we assume a cost of US \$5,000, that would imply a cost of US \$175M to store one Exabyte of data. If this price is correct, this is already a significant reduction in cost compared to the above examples of cloud-based storage.

What becomes more complex, and already an issue, are the security and access rights to the data that is stored. This ranges from the rights of individuals themselves, to their DNA data, to the rights of governments, insurers, health providers, law enforcement and many more organizations and areas that could find uses for this data. In other words, the data sharing. While raw storage costs are coming down, the additional layers that will need to be tacked on top of this for data sharing may be where the bottlenecks emerge, particularly as the industry grapples with the ethical and moral issues surrounding the access and use of this data.



Samsung 30TB Hard Drive

# Data Sharing

Storage of data in and of itself for genomics, aside from cost management, is not really a critical issue in isolation. What becomes more complex is how big data is shared and subsequently processed and analyzed in addition to security and privacy concerns. Of the other areas dealing with big data, probably only astronomy comes closest to genomics in terms of the volume of data to be dealt with. However, in contrast to genomics, most of the data acquired and used in astronomy is analyzed and dealt with in close proximity to the telescopic equipment that acquired the data. There are some astronomy projects like the Square Kilometre Array (SKA) which requires data from 3,000 distributed antennae to be sent to a central server. This project apparently requires up to 600 terabytes/second of bandwidth according to Zachary D. Stephens et al., “Big Data: Astronomical or Genomical,” PLOS Biology 13, no. 7 (2015). That type of bandwidth just does not exist, particularly across the commercial internet. While on an aggregate level, commercial services like Netflix, Amazon Prime and Youtube use a lot of bandwidth the typical consumer connection of 10Mbps is sufficient to support the individual data streams.



The Square Kilometre Array

In the US, in an increasing number of towns and cities, it is now possible to get fiber connection to a home delivering 1 Gigabit per second (1000 Mbps) internet connection for a cost ranging from \$70 to \$120 per month. For comparison, a mobile network carrier is able to deliver 5 to 12 Mbps to a smartphone via 4G and also charges consumers approximately \$70 to \$120 per month.

Besides the direct cost in money, the big cost in data transfer is time. It is estimated that downloading 1 TB of data with a transfer speed of 10 Mbps over Thin Ethernet would take over 10 days. With 100 Mbps Fast Ethernet it would take slightly over a day, and with one of the fastest available speeds today, a 1000 Mbps GIG E transfer speed, it would take about 2 and a half hours. If 1 Petabyte of data had to be transferred, with 1000 Mbs GIG E transfer speed, it would take over 104 days. For any use case which would require collecting big data for analysis from different sources, this presents a daunting challenge. Genomics data is distributed across centers in various countries and labs and also spans a wide range of sizes. For data sets approaching 1 Petabyte or greater, physical transport of data on hard disks or other media would probably be quicker in these cases than attempting to transfer it over telecommunications networks.



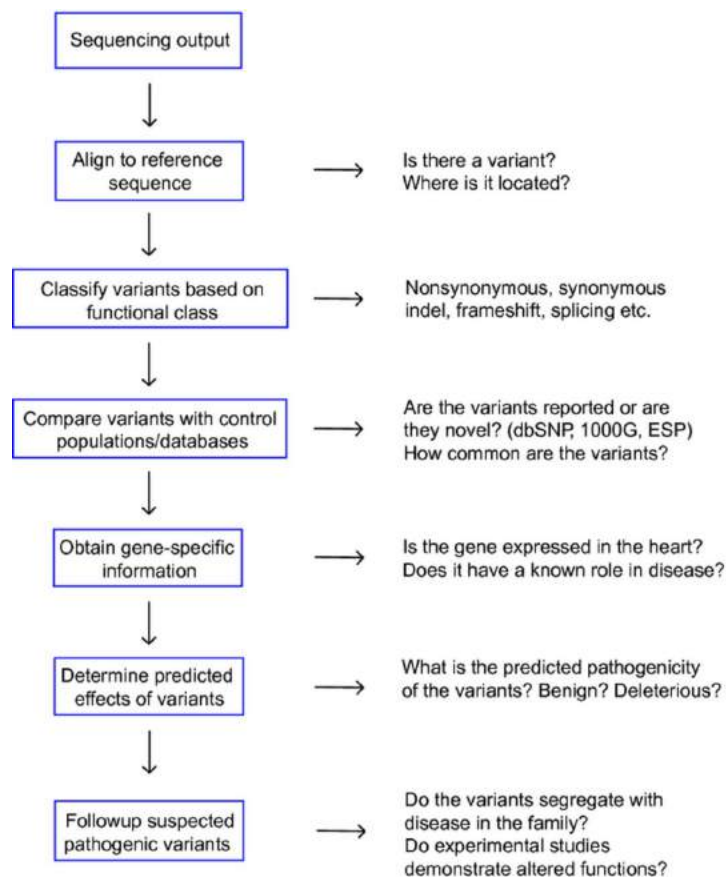
Moving small amounts of data is pretty straightforward across the public Internet. Moving larger amounts of data, Big Data, requires more detailed planning. Compression tools may be one area that helps improve the efficiency of transfer techniques. Hardware and technical limitations may in some cases constrain the amount of research that can be done. Research or experiments that may take a few months to just get the data may receive lower priority than other projects that do not have such constraints.



# Data Processing and Analysis

Once the data required for being processed and analyzed has been collected and formatted, this is where a significant challenge of the big data created by genomics comes into play. Currently processing and analysis is quite intensive, requiring complex algorithms and many sets of computational instructions to be executed.

Flow chart showing steps for DNA sequence analysis. ESP indicates Exome Sequencing Project; 1000G, 1000 Genomes project.



Up until early June 2018, the Chinese built - Sunway TaihuLight, held the position as the world's fastest functioning super computer. It is located in the Jiangsu province of China and can currently process 93 petaflops. One petaflop is a quadrillion floating-point operations per second. One quadrillion is a thousand trillion or  $10^{15}$ . It is estimated to have cost US\$ 273M to build. The Chinese are using it for climate modeling, life science research, advanced manufacturing and various data analytics. In June 2018 it was reported that the US, who had been falling behind in fast computing, have finally completed work on the previously announced super computer, named the Summit, at the Oak Ridge National Laboratory. Summit boasts a capacity of almost 200 quadrillion operations per second and is estimated to have cost about \$102M to build. These computers are well suited for processing big data, however, the use and accessibility of these machines are still limited.

Despite the limited availability of super computers with massive processing capability, the cost and availability of powerful processors are becoming accessible in a more broad and wider manner. In June of 2017, Intel released their Core i9 processor which goes for about US\$ 2,000 and can execute at a teraflop, which is one trillion floating-point operations per second.

As with data transfer, the processing power available to a particular facility may dictate constraints around what experiments, analytics and development can be accomplished.

# Cloud-based Genome Data Storage and Application Providers

|                       | Offering   | Data storage | Secondary analysis | Tertiary Analysis | Data infrastructure                                |
|-----------------------|--|--------------|--------------------|-------------------|--|
| Google Genomics       | Securely store, process, explore, and share large, complex biological datasets.  | ✓            | ✓                  | ✓                 | Google Cloud                                       |
| Microsoft Genomics    | Powers genome sequencing and research insights   | ✓            | ✓                  | ✓                 | Micorsoft Azure                                    |
| IBM BlueBee           | High-speed analysis in secured data centers that meet or exceed organizational and geographical data protection requirements         | ✓            | ✓                  | ✓                 | IBM Cloud  |
| DNAexus               | Makes large genomic datasets broadly accessible to the research community, coupling them with analysis tools                         | ✓            | ✓                  | ✓                 | Microsoft Azure, Google Cloud, Amazon Web Services |
| BaseSpace by Illumina | Data flows directly from sequencer into BaseSpace, enabling data management and analysis using a curated set of analysis apps        | ✓            | ✓                  | ✓                 | Amazon Web Services                                |
| DNASTack              | Cloud-based genomics platform that helps geneticists securely manage, analyze, search, and share genomic datasets in the cloud       | ✓            | ✓                  | ✓                 | Google Genomics                                    |
| Qiagen CLC            | Facilitates analysis when the same analysis needs to be performed repetitively, for example, on multiple samples                     | ✓            | ✓                  |                   | Amazon Web Services                                |
| Seven Bridges         | Cloud-based environment for conducting bioinformatic analyses. Store, analyze, and jointly interpret bioinformatic data              | ✓            | ✓                  | ✓                 | Private cloud, Amazon Web Services, Google Cloud   |
| Curoverse             | Proves open source enabling technology to address data and compute infrastructure software challenges in the biomedical field        | ✓            | ✓                  | ✓                 | Private Cloud, Curoverse cloud                     |
| Golden Helix          | Provide a complete end-to-end solution for clinical labs and hospitals to analyze Next-Generation Sequencing data                    | ✓            | ✓                  | ✓                 | Golden Helix Cloud                                 |
| Fabric Genomics       | Seamless, fully integrated NGS workflow from raw data to clinical report   | ✓            | ✓                  | ✓                 | Fabric Genomics Cloud                              |
| Syapse                | Works with health systems to implement precision medicine programs , enabling oncologists to deliver personalized care               | ✓            |                    | ✓                 | -  |
| Luna DNA              | Community owned database that rewards individual shares in the database and rewards for contributing their DNA                       | ✓            |                    |                   | Blockchain   |
| DNA Land              | University owned database. Donate genome data to enable scientists to make discoveries for the benefit of humanity.                  | ✓            |                    |                   | DNA land cloud                                     |
| Nebula Genomics       | Leverages blockchain technology to eliminate the middleman and empower people to own their personal genomic data.                    | ✓            |                    |                   | Blockchain   |
| Sophia Genetis        | Platform for performing diagnostic testing: facilitates visualization and interpretation of variants while assuring data protection. |              | ✓                  | ✓                 | -  |
| Spiral Genetics       | Large-scale DNA data analysis for medical, pharmaceutical, and agricultural research.  |              | ✓                  | ✓                 | -  |
| Genomenon             | Curated database of disease-gene-variant relationships found in the full text of the scientific literature                           |              |                    | ✓                 | -  |
| SolveBio              | Aggregates genetic datasets that researchers can efficiently analyze to better diagnose diseases. "Bloomberg" of genetics            |              |                    | ✓                 | -  |
| Diploid               | Machine learning diseases variant interpretation engine  |              |                    | ✓                 | -  |

Source: GenomSys analysis. Vendor websites.



"Genomic data  
is one of the  
fastest growing  
datasets in the  
world."

MICHAEL J. MCMANUS, PHD  
SR. HEALTH & LIFE SCIENCES SOLUTION ARCHITECT AT INTEL CORP.



# DEVELOPMENT OF THE MPEG-G STANDARD

In its 30 years of activity Working Group 11 of ISO/IEC Joint Technical Committee 1 Sub Committee 29 – also known as Moving Picture Experts Group (MPEG) — has developed many generations of successful standards that have transformed the world of media from analog to digital.

In 2016, MPEG and ISO TC 276 began working to produce MPEG-G, a new open standard to compress, store, transmit and process genome sequencing data. The standard plans to reach a compression factor for raw data of approximately 100 which means an improvement of up to one order of magnitude with respect to currently used formats. MPEG-G will provide new functionalities such as native support for sophisticated selective access, hooks to implement any data protection mechanism, flexible storage and streaming capabilities. This will enable various new applications, such as real-time streaming of data from a sequencing machine to remote analysis centers during the sequencing and alignment processes.

MPEG-G plans to be finalized sometime in 2018 and officially published in 2019. During the transition from existing formats to MPEG-G, interoperability and integration with existing genomic information processing pipelines will be provided by interfaces and transcoders from/to the legacy file formats (FASTQ/SAM/BAM/CRAM).

# MPEG-G

## Compression Performance

In order to assess the compression performance of genomic data file formats, the notion of “coverage” (or “depth”) has to be introduced. The term coverage in genomic sequencing identifies the number of unique sequenced fragments that include a given nucleotide in the reconstructed sequence. The higher the coverage, the higher the confidence that the reconstructed sequence is actually the one that has been sequenced by a sequencing device. High values of coverage (usually larger than 50) are required for clinical applications.

Currently the most used file format for compressed genomic data is BAM, which has been recently improved (between 30% and 50%) by CRAM. It is important to stress here that while BAM and CRAM enforce both the encoding and decoding process by fixing the compression algorithm both at the encoder and decoder side, MPEG-G only standardizes a bit stream syntax and a decoding process. This approach leaves the field open to competition among developers aiming at implementing more and more efficient encoders as long as the generated syntax is compliant with the standard specification.

As of this writing, tests on MPEG-G compression performance are still ongoing, but the first figures show an improvement of a factor 2 to 3 with respect to gzip (the current de-facto standard for raw data compression) and between 30% and 50% with respect to CRAM.

\*For more information: <https://mpeg.chiariglione.org/standards/mpeg-g>

MPEG-G

# MPEG-G MEETS GENOMICS

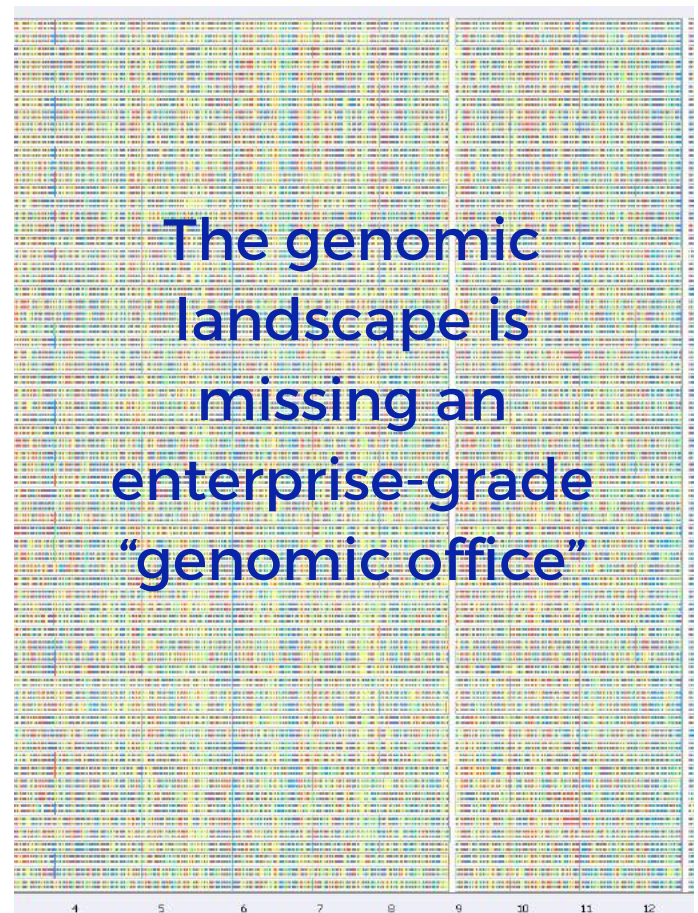
## Unique opportunities for first movers

The history of digital media (from mp3 to Ultra-HD TV) has shown that a new ISO standard for the representation of digital information represents a unique opportunity for the creation of an entirely new market of tools and technology. The standardization of digital information representation enables content creators to reach a huge amount of content consumers thanks to the availability of a competitive marketplace of content processing and sharing devices. Today MPEG-G represents this very opportunity for a genomics data processing market which is still in its infancy partly due to the lack of such standard framework. This situation is a unique opportunity for first movers who can get to the market the first products with standardized interfaces able to read and write content in conformity with the new standard. While the genomic market analysis is rather fragmented according to the wide range of existing applications and fields of research, a few core applications are needed horizontally by all players to perform basic actions on the genomic content.

## Editing Tools

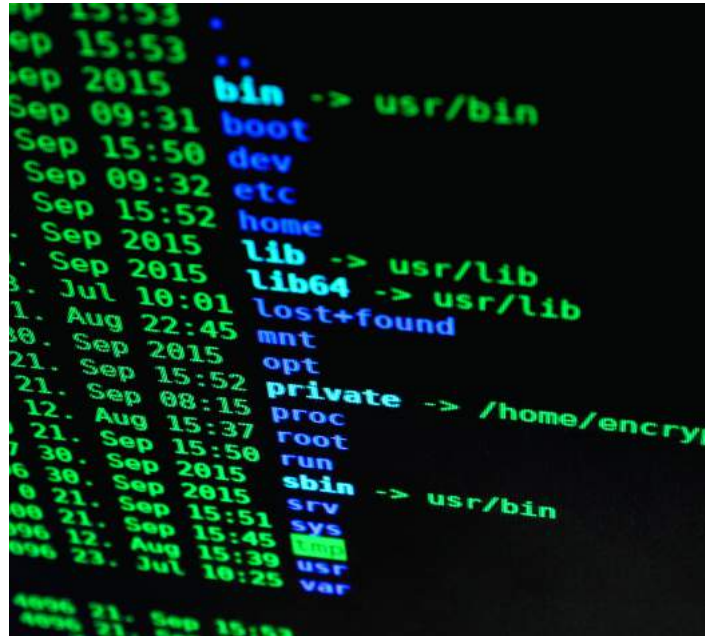
---

Editing tools are software applications enabling the manipulation of digital content. In case of MPEG-G an editing tool enables reading and writing MPEG-G compliant files by means of decompressors and compressors. An editing tool usually provides a graphical user interface (GUI) to select, browse and modify content using graphical indicators such as buttons and icons. Additional features may include the capability to retrieve remote content or send local content to remote locations via streaming or upload/download. A typical example of an editing tool is Microsoft Word for textual documents. In this case the (de-facto) standard is the .doc file format specification and the most pervasive editing tools is Microsoft Word followed by several other minor competitors (Open Office, etc.). For now the genomic landscape is missing an enterprise-grade “genomic office,” but things may change as soon as a serious development effort will be put in the implementation of an MPEG-G editing tool.



# Security and Privacy Protection

In order to enable large-scale genomics data processing applications, policies are being put in place to guide data sharing in a way that protects and promotes the confidentiality, integrity, and availability of data and services, and the privacy of individuals whose genetic profiles are shared. In order to preserve data security and personal privacy, adequate mechanisms of anonymization, access control, integrity check, traceability, role segregation shall be implemented at all stages of the genomic analysis pipelines.



This need will become even more urgent when the standardization efforts of MPEG and other bodies such as the Global Alliance for Genomics and Health (GA4GH) will achieve their goals of making genomics medicine common practice at all levels of public health care. This situation creates a great opportunity to develop security applications on top of genomic data representation formats such as MPEG-G which supports finely tuned access control schemes and privacy rules hierarchies. Such applications will become mandatory when governments and regulators will define the type of access control requirements for large-scale genomics medicine projects. A genomic security application is essentially an editing tool with extensions to apply encryption, digital signatures, time-stamping, privacy rules definition and any other mechanism to govern access to the encoded genomic data.

# Genomic Analysis Tools



Companies currently marketing applications handling genomic data in the currently adopted formats can seize the opportunity to be the first movers in adopting MPEG-G and therefore offer their existing customers new options of data compression, transport and storage. This would give first movers an edge over competitors and maybe enable runner-ups to become market leaders in a market, which is still extremely fluid with undefined boundaries between leaders and underdogs.

Today when you stream video on YouTube or Netflix you don't really know (and care) what's the underlying file format, but only care about the quality of the image and the number of features offered to interact with it. Since MPEG-G promises to offer not only better compression but a wide set of new options to interact with content, the integration of an existing application with MPEG-G enabled interfaces could really generate a shift in the customers' perception and decisions.



# The Data Focused

These 4 under the radar companies are out to address and solve the genomics data compression and data transfer challenges.



GenomSys is located in Lausanne, Switzerland. They are among the first developers of MPEG-G based solutions to enable genomic applications with advanced features of data access and handling as well as with a dramatic reduction of both storage costs and transfer time from sequencing facilities to storage and/or analysis sites. They have to date received US\$ 2 million in funding.

<http://genomsys.com>



Located in Sunnyvale, US, Geneformics creates products that employ compression and other cutting-edge technologies to streamline the storage and sharing of genomics data, on-premises and in the cloud. They have received total funding to date of US\$ 2.9 million.

<http://www.geneformics.com>

\*as of this writing/publication the link to the Geneformics website is currently not active



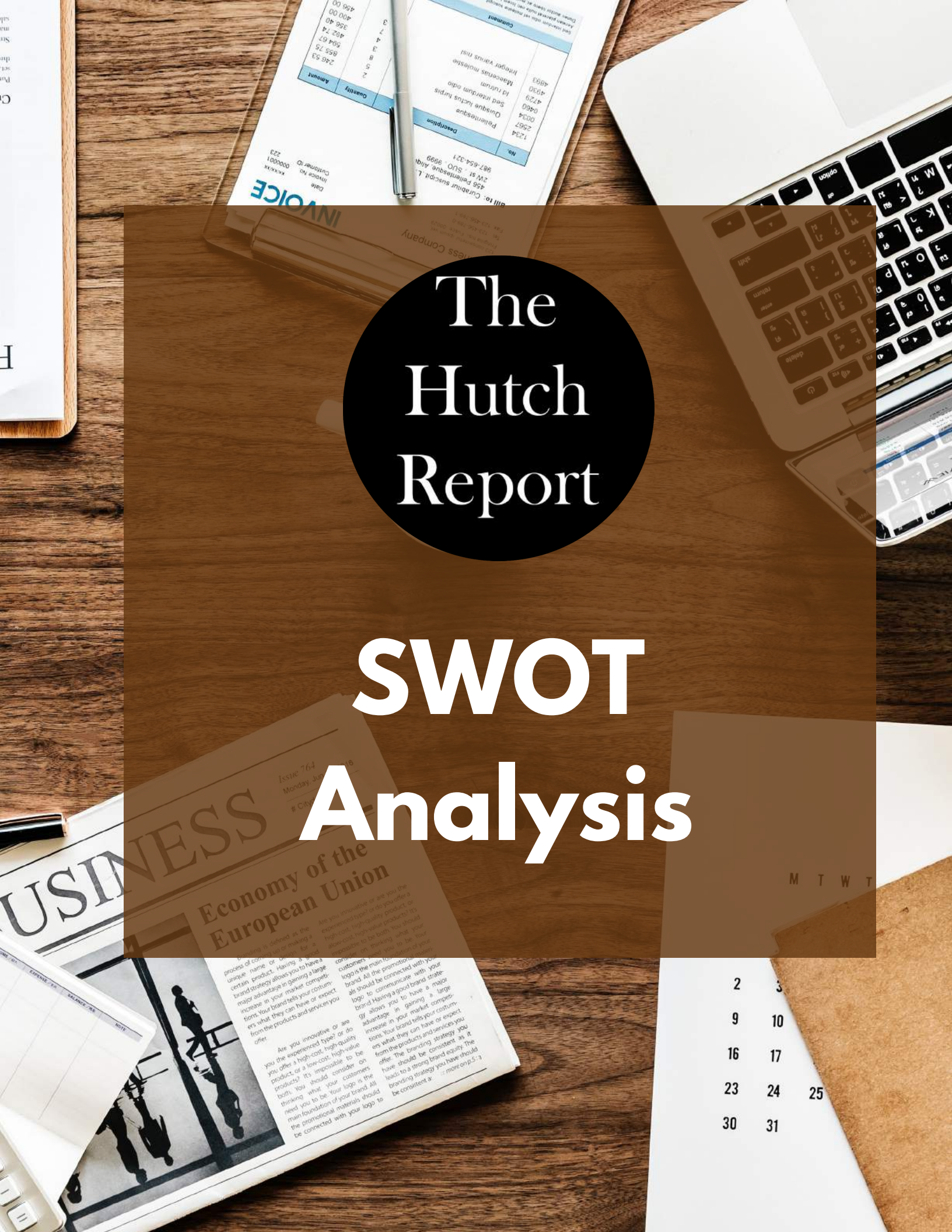
PetaGene is a bioinformatics company that has created a novel, best-in-class method to compress and manage genomic data. The company won first place at BioIT World (the top conference for the industry). Located in Cambridge, UK, it was founded by a University of Cambridge PhD. They received roughly US\$ 251k in funding April 6, 2017.

<https://www.petagene.com>



Annai Systems is a leading genomic data analysis company committed to developing cost-reducing innovations that accelerate scientific discovery, maximize statistical power, and optimize data control. The company offers solutions from data compression and storage to secondary and tertiary analysis. They have to date received US\$ 13.6 million in funding.

<http://annaisystems.com>

A top-down view of a wooden desk. In the upper right, a silver laptop is partially visible. In the upper center, a pen rests on an invoice. The invoice has a table with columns for 'Description', 'Quantity', and 'Amount'. Below the table, it says 'INVOICE' and 'Business Company'. In the lower left, a newspaper is open, showing the word 'BUSINESS' and an article titled 'Economy of the European Union'. In the lower right, a calendar page shows dates from 2 to 31. A large, semi-transparent brown circle is centered over the desk, containing the text 'The Hutch Report'.

# The Hutch Report

# SWOT Analysis

## ▶ **Minimal Physical Risks**

The actual physical risks associated with most genetic tests are very small, particularly for those tests that require only a blood sample or buccal smear (a method that samples cells from the inside surface of the cheek).

## ▶ **Benefits to Family Members**

Because genetic conditions often run in families, information about your genetic makeup might be useful to other family members. If family members are aware that a genetic condition runs in the family, it might prevent them from being misdiagnosed. This information might also be of use to them when they are planning children.

## ▶ **Avoiding Unnecessary Investigations**

For many disorders, genetic testing is the best way to make an accurate diagnosis. An accurate diagnosis can then guide the clinician in choosing the most suitable therapy and support for the patient. Therefore, a negative result can eliminate the need for unnecessary checkups and screening tests in some cases. For example, genotyping cancer cells and understanding what genes are misregulated allows physicians to select the best chemotherapy and potentially expose the patient to less toxic treatment since the therapy is tailored.

## ▶ **Sequencing Costs**

Our ability to analyze the terabyte of data generated by sequencing one genome is improving, as dozens of big data startups and a torrent of venture capital pour into the hot new genome interpretation space making it cheaper and cheaper to accomplish. The most recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionised the study of genomics and molecular biology.

## ▶ **Emotional Relief**

A definite diagnosis through genetic testing can provide a sense of relief to patients and families from uncertainty, especially if they have been searching for the answer for long time. This in turn allows them to make better informed decisions about managing their health care.

## ▶ **Genetics vs. Lifestyle**

The truth is that in spite of a person's genetic makeup, people in general are not living healthy lives. No amount of genetic data will ever change the fact that most of us need to eat better and move more. Relying on understanding your genetic make up as a cure all solution will divert attention from many of the small improvement that each of us can make daily.

## ▶ **Requires Substantial Investment**

Genome sequencing costs may be coming down but this is still a very complex technology. Next-generation sequencing requires sophisticated bioinformatics systems, fast data processing and large data storage capabilities, which can be very costly.

## ▶ **Necessary Resources**

Although many institutions may have ability to purchase next-generation sequencing equipment, many lack the computational resources and staffing to analyse and clinically interpret the data. Most physicians are not trained in how to interpret genomic data.

## ▶ **More Noise than Signal**

One of the main problems with genetics is that the research is bound to produce more noise than signal, however, the issue isn't genetics per se but "big" data. The basic idea of precision medicine is to look for patterns in the genome that seem to travel with problems we all care about: diabetes, heart disease, cancer and dementia. But there are a lot of possible patterns to look for. In other words, there are too many variables and too many different combinations of genomic and clinical data to detect meaningful patterns. We have accumulated huge databases on human genetic differences — but many of the differences appear to be more or less irrelevant.

## ▶ **Margins of Error**

It is important to understand that there are no real certainties in genetic testing. Genetic testing only indicates probabilities, which are not 100 percent certain. Despite breakthrough advances in DNA genetic testing, gene tests simply cannot rule out every possibility of developing a disease. Positive tests also do not necessarily mean a patient will necessarily develop a disease or disorder soon, if ever. Thus, there are certain limitations on scientific research technology and the margin of error of certain tests.

## ▶ **Data Quality**

Researchers are currently facing substantial challenges in storing, managing, manipulating, analyzing, and interpreting whole-genome sequencing data even for just moderate numbers of individuals as they need to take into account the data quality of information stored in BAM files. These challenges will become exacerbated when millions of individuals are sequenced. Adding to that, the role of most of the genes in the human genome is still unknown or incompletely understood. Therefore, a lot of the "information" found in a human genome sequence is unusable at present.

## ▶ **Inefficient Data Sharing Methods**

Data sharing is inefficient and compromises security. Data is shared between organizations over file-based transfer protocols (e.g., FTP), requiring large volumes of data to be downloaded over the internet onto a local hard drive before it can be used, which propagates data redundantly and consumes expensive network bandwidth and data storage capacity. By repeatedly transferring data out of a single secured environment, stewards of sensitive, identifiable genomics information lose the ability to oversee access to bona fide researchers or audit their use of it.

## ▶ **Insufficient Data**

There is a huge gap between data that is publicly available and the type of information that is needed for diagnostic decisions.

## ▶ **Income Generating Potential**

The Genomics market is expected to reach US\$ 23.88 Billion by 2022 from an estimated US\$ 14.71 Billion in 2017, at a CAGR of 10.2%. The key factors such as growth in the number of R&D activities, growing demand for personalized medicine and reducing cost of sequencing services are driving the growth of this market.

## ▶ **Healthcare Innovations**

The application of knowledge gained from the characterisation of the genomes of several organisms, including the human genome, holds considerable potential for the development of new health care innovations over the coming decades. We are able to assemble enough information to study particular conditions at an ever-increasing rate. We are discovering, categorizing and exploring treatments for conditions that are considerably narrow in definition. The power of precision medicine to treat and even heal is unprecedented.

## ▶ **New Insights**

Work in the field of genomics will also offer completely new insights into the mechanisms of human and animal development and ageing; and, because our evolutionary history is written in our DNA, it will start to unravel our genetic roots and help us to understand the relationships between and within different species.

## ▶ **Prevention and Diagnosis of Genetic Conditions**

It is now believed that the information generated by genomics will, in the long-term, have major benefits for the prevention, diagnosis and management of many diseases which until now have been difficult or impossible to control. If a patient is provided with an accurate diagnosis, the appropriate treatment can be given. For example, if a genetic test tells you that you have an increased risk of developing a condition later in life (such as breast cancer) you can take measures to keep the risk to a minimum, such as going for more regular check-ups. In a 2013 article for the New York Times, Angelina Jolie explained her decision to undergo a double mastectomy after finding that she had a mutation in a gene known as BRCA1 that greatly increased her risk of breast and ovarian cancers.

## ▶ **Drug Development and Increased Efficacy**

Research directed at pathogen genomes will enhance our understanding of disease transmission and of virulence mechanisms and how infective agents avoid host defences, information which should enable the development of new classes of diagnostics, vaccines and therapeutic agents. Whole human genome sequencing will therefore help scientists trace and design effective drugs by evaluating the patient's genes to discover variants. In other words, genome study can enhance the efficacy of drugs which are otherwise ineffective for certain patients due to their intrinsic genes.

## ▶ **Detecting Abnormalities**

The most advanced methods can now detect abnormalities across the entire genome (whole-genome sequencing only), including substitutions, deletions, insertions, duplications, copy number changes (gene and exon) and chromosome inversions/translocations. A major strength of next-generation sequencing is that it can detect all of those abnormalities using less DNA than required for traditional DNA sequencing approaches.

## ▶ **Accentuated Healthcare Inequalities**

Because many of the medical benefits of genomics research may, at least at first, be very expensive, there is a danger that these new developments will increase the disparity in health care within and between countries. There are particular concerns that inequalities in health care will be accentuated by the current trends in the management of intellectual property, particularly the patenting of basic genomic information.

## ▶ **Knowledge Management**

The information we obtain from genetic testing becomes more valuable the more available it is for others to analyse. Lack of sharing, or proper knowledge management of this information will impede the opportunity to discover more. In a recent publication of Nature the American College of Medical Genetics and Genomics had this to say on the topic:

“The considerable variation in clinical presentation and molecular etiology (the science of finding causes and origins) of genetic disorders, coupled with their relative individual rarity, makes it clear that no single provider, laboratory, medical center, state or even individual country will typically possess sufficient knowledge to deliver the best care for patients in need of care.”

## ▶ **Poor Diagnosis**

With the costs of genome sequencing coming down, the number of labs that offer gene sequencing at cut-rate prices will most likely increase. The data obtained from these labs will most likely scare many — particularly since they are likely to be framed as a 50% increase in their risk of Disease X. However, it's just as likely they won't make a difference to your health.

## ▶ **Privacy**

Legitimate concerns have been raised about giving your DNA to testing companies, who then legally own the data and can share it with multiple entities beyond your control. For example, consumers run the risk of government agencies and other agencies using their genetics information against them by 1) Sharing it with other organizations for research purposes; 2) Denying insurance on the basis that their genes make them liable to suffer from grave diseases; and 3) Dictate medical policies using unfair practices.

## ▶ **Scope for financial exploitation**

The sheer commercial parlance of genome sequencing leaves a lot of financial manipulation by pharmaceutical, federal and scientific bodies that are solely dictated by commercial motives. This is especially true in the case of customized drug manufacturing which can cause ambivalence to many. In addition, no healthcare system or pharmaceutical company will desire the interpretation of data unless they know it to have a relevant application.

## ▶ **Controversial**

Genomics is inherently a politically divisive topic that polarizes various groups/parties. Religious and other related organizations argue that whole human genome sequencing is a sophisticated strategy to tamper and interfere with the laws of the Divine and also nature. However, other progressive entities and scientists feel otherwise. The number of ethical and legal issues regarding the field of genomics cannot be understated.

## ▶ **Data Policies**

An individual's genome may contain information that they DON'T want to know. For example, imagine a patient has genome sequencing performed in order to determine the most effective treatment plan for high cholesterol. In the process, researchers discover an unrelated variant form of a gene (allele) that assures a terminal disease with no effective treatment. Without very strict data policies it may be hard to determine what to disclose or not to disclose to the patient.

## ▶ **Data is not democratized**

Access to data and infrastructure isn't democratized. Centralization of data and infrastructure in a small number of institutions restricts access by the wider community, and its growth. Institutional authorization controls often hinder access by external users or preclude it altogether. Collaborative projects usually bring together researchers who are not affiliated with a single institution and who may not be traditional scientists (e.g., researchers from clinical diagnostic laboratories, industry, other fields of study, or citizen scientists).

## ▶ **Methods are not reproducible**

Data processing workflows are developed in-house and hardcoded to run on specific local hardware architectures, limiting their reproducibility and confounding the comparison of results. The resulting data are not easily comparable, limiting the reuse of valuable information and delaying their interpretation and applications to human health.

# Genomics and the Blockchain

As we have illustrated throughout this report, there is a bottleneck in genomics and it comes in the way of data, lots of data. We have highlighted the need for better storage capacity, data cohesiveness, security, standardised delivery mechanisms, and more precise analysis of data if the field of genomics is going to advance and deliver all that has been promised.

One of the current go-to technologies that could provide a solution to genomics digital privacy issues, and delivering secure data sets without having to trust third-parties is the Blockchain. Already different companies are betting on the need, based on the expected phenomenal growth of genomic data, for blockchain-based securitization of data. Besides some hype about the connection of blockchain and genomic data, viable business models and how to integrate the technology into the existing healthcare industry are far from clear.

Data privacy and secure handling of data are a major concern for businesses and private genome owners. Genomic data are very sensitive to many kinds of possible breaches and exploitations, such as higher health insurance fees based on chronic diseases revealed by personal genome data or the unauthorized use of genomic data for commercial drug development.



Together, blockchain and decentralized storage could be a promising technology for managing and securing the coming genomic data deluge. One way how this could work out is the following: sequenced genomes, transcriptomes or proteomes from different sources, research institutes, hospitals, sequencing companies or universities, are stored on a decentralized infrastructure like IPFS. Every personal genome file is then tagged or identified with a unique unhackable hash number for transactional traceability on the blockchain. The genomic data are not stored on the blockchain, but on a decentralized infrastructure, where no central institution or company has control over them. Only transactions in the broadest sense between different parties are then stored on the digital ledger of the blockchain.

If, for example, a research institute requires genomic data with certain properties for understanding the genetic conditions of a disease, they could issue a request to the genome holder who has stored his/her genome on the decentralized blockchain infrastructure. As in the case of current Bitcoin transactions, the exchange of data between the genome owner and the research institute would then be recorded on the blockchain, where it is possible to trace the use of the data later on. This exchange could be paid for by the institute, opening reimbursement opportunities for the genome holder.

There are still a large number of challenges with current genomic data being widely disseminated between a plethora of institutions from hospitals, universities, research institutes, sequencing and analytics companies to private storage.

In addition, blockchain technology is still going through its own growing pains and is not without its share of challenges before hoping to become widely accepted.

Therefore the big challenge will be to fit the blockchain technology into the genomic and health landscape in general. Current business models of genomic blockchain companies are mostly similar, with slight variations. Only time will tell which business model can be expanded and refined until it is most suited to establish the blockchain as a technology to support our genomic future for the benefit of everyone.

In spite of these challenges, it has not stopped businesses from trying to develop in this direction as investment continues to flow into this space. Following are the principle companies active in the cross section of genomics and blockchain technology.



# Genomics / Blockchain Startups

## **EncrypGen**

Website: [www.encyrpgen.com](http://www.encyrpgen.com)

Funding Stage: Seed

EncrypGen provides customers and partners best-in-class, next generation, blockchain security for protecting, sharing and re-marketing genomic data. To date they have received US\$ 500,000 in seed funding.

## **Shivom**

Website: [www.shivom.io](http://www.shivom.io)

Funding Stage: ICO

Shivom combines blockchain, AI, DNA sequencing & cryptography to enable secure and personalized medicine. The Shivom platform works on principles of collaboration & integrity, allowing people to own, manage and monetize their data. By creating a web-marketplace, a network of genomic counsellors, and a not-for-profit drug research unit, Shivom will build a global healthcare ecosystem, reaching even low-income countries where such services have not been previously available. They have received US\$ 35 million from 5 investors and launched an ICO on May 21, 2018.

## **Genomes**

Website: [www.genomes.io](http://www.genomes.io)

Funding Stage: Unknown

Genomes aims to sequence 1 billion people's genome and power new insights by empowering the owners to control access to their genome. Users can grant selective, controlled, audit-able access in exchange for financial reward (OME tokens) or to ask questions of their data as the field of genomics discovers more about our genetic code. Genomes uses <https://rockchain.org/> technology to provide private querying of personal genomes on the Ethereum blockchain.

**Zenome**

Website: [www.zenome.io](http://www.zenome.io)

Funding Stage: ICO

The Zenome is a non-profit organization that aims at raising awareness on genomic medicine and providing the platform for DNA exchange. The most fundamental duty of Zenome is to ensure ownership of personal genomic information and equal access to the market of genomic information. Being a distributed genomic market, based on P2P network and blockchain, Zenome puts users in control of their genetic data. The fairness of data exchanges is enforced by smart-contracts and utilization of dedicated tokens (ZNA). Together with genetic services of any kind the platform is expected to provide one-stop ecosystem for genetic data. The ICO generated roughly US\$ 600,000 in funding.

**BlockGene**

Website: [www.blockgene.io](http://www.blockgene.io)

Funding Stage: Unknown

BlockGene presents themselves as a blockchain-based genomic trust system that includes integration, authorization, search, and provision. Blockgene.io integrates into existing ecosystems to provide more genomic data for drug development.

**LunaDNA**

Website: [www.lunadna.com](http://www.lunadna.com)

Funding Stage: Series A

Luna DNA, is a genomic and medical research database powered by the blockchain. Created by co-founders of the \$40B DNA sequencing leader Illumina, Luna DNA incentivizes the sharing of DNA and medical information for research. Luna rewards people for sharing the data they already own while contributing to medical research and discovery that saves lives. LunaDNA has received US\$ 4 million from 8 investors in 2 funding rounds.

**E-Nome**

Website: <https://www.enome.io>

Funding Stage: Unknown

E-Nome is an Australian technology startup driving the application of blockchain technology to the secure storage of health records. The Garvan Institute of Medical Research (Sydney) has signed a memorandum of understanding with E-Nome to assess the potential of the E-Nome platform for the secure storage of genomic information.

## **Nebula Genomics**

Website: [www.nebulagenomics.io](http://www.nebulagenomics.io)

Funding Stage: ICO

Nebula Genomics was started by pioneering Harvard and MIT geneticist George Church. They leverage blockchain technology to eliminate middlemen such as 23andMe and AncestryDNA, and empower people to own their personal genomic data. This process is expected to effectively lower sequencing costs and enhance data privacy, resulting in growth of genomic data. Its open protocol will leverage the genomic data growth by enabling data buyers to efficiently aggregate standardized data from many individual people and genomic databanks. Nebula ended an ICO token sale in April 2018 which raised roughly US\$ 5.8 million.

## **Longgenesis**

Website: <http://longgenesis.com>

Funding Stage: ICO

Based in Hong Kong, Longgenesis is a revolutionary blockchain-based Life Data Marketplace platform which provides modular toolsets coupled with integrated advanced Artificial Intelligence (AI) systems to store, manage, and trade life data. Essentially, Longgenesis is a life data marketplace that aims to facilitate the sale/purchase of human data between Users (General public) and Customers (Drug development/Pharmaceutical companies). Longgenesis is a joint venture by Bitfury and Insilico Medicine, two leaders within the Blockchain and Medical AI industries. LifePound is the Cryptocurrency launched as the central monetary tool of the Longgenesis Marketplace. They recently announced a partnership with Nebula Genomics.

## **X Genomics**

Website: <https://www.xgenomics.org>

Funding Stage: ICO

X Genomics is a project combining gene technology with blockchain technology. It will break the silos of industry information and integrate the human genome data on a blockchain by building a global genetic data hub and opening a transparent service platform for sharing human genome data. The team of researchers and genome experts supporting the program is based in Singapore, Canada and China, and include two Nobel prize winners, Professor Randy Schekman, 2013 Nobel Prize in Physiology or Medicine and Eric Maskin, Harvard Professor and 2007 Nobel Prize in Economy. This makes X Genomics the first of its kind to be supported by Nobel Prize laureates. To power its blockchain, The X Genomics project will issue an ERC-20 token, the X Genomics Chain Token, (symbol: GSX) under the Human Genomic Research Foundation LTD, founded in Singapore.

**Doc.AI**

Website: <https://www.doc.ai>

Funding Stage: ICO

Doc.AI has a mission to decentralize precision medicine on the blockchain. They believe that the biological profile of the near-future will be consumer-controlled, blockchain-based, AI-powered and OMICS-data-centric. The Neuron (NRN) tokens give access to the AI network and reward the users with tokens for training their AI. With these tokens they can broadcast a competition on NEURON and create a bounty (prize) for data scientists.

**YouBase**

Website: <https://www.youbase.io>

Funding Stage: ICO

Headquartered in Englewood, CO, YouBase combines blockchain compatible technologies which together deliver a secure and flexible container for data that is independent of any one single entity. YouBase is paving the way for a future-proof approach to storage & management of data around individuals.

# General References and Sources

Eleanor Rieffel, Quantum Computing: A Gentle Introduction,

<http://mmrc.amss.cas.cn/tlb/201702/W020170224608150244118.pdf> , 2011

Fabrcio F. Costa, Ph.D., Big Data in Genomics: Challenges and Solutions,

<https://pdfs.semanticscholar.org/0e38/7bd00952c8450deefcdbcdeb5c946c20f54.pdf>

Karen Y. He, Dongliang Ge, and Max M. He, Big Data Analytics for Genomic Medicine,

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343946/>

Stonefly, Big Data Storage Challenges and Solutions, <https://stonefly.com/blog/big-data-storage-challenges-solutions-genomics>

Zachary D. Stephens, Skylar Y. Lee, Big Data: Astronomical or Genomical?

<http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195>

Bonnie Feldman, Ellen M. Martin, Tobi Skotnes, Big Data in Healthcare Hype and Hope,

[https://www.ghdonline.org/uploads/big-data-in-healthcare\\_B\\_Kaplan\\_2012.pdf](https://www.ghdonline.org/uploads/big-data-in-healthcare_B_Kaplan_2012.pdf)

Meredith Salisbury, Challenges For Genomics In The Age Of Big Data,

<https://www.forbes.com/sites/techonomy/2015/07/23/challenges-for-genomics-in-the-age-of-big-data/#5e1bba0d2072>

Wikipedia, Compression of Genomic Re-Sequencing Data,

[https://en.wikipedia.org/wiki/Compression\\_of\\_Genomic\\_Re-Sequencing\\_Data](https://en.wikipedia.org/wiki/Compression_of_Genomic_Re-Sequencing_Data)

Sebastian Deorowicz, GDC 2: Compression of large collections of genomes,

<https://www.nature.com/articles/srep11565>

Ka-Chun Wong, Big Data Challenges in Genome Informatics, <https://arxiv.org/abs/1803.09632>

Sai Venkatesh Balasubramanian, Genome Data Compression using Digital Chaos, <http://vixra.org/pdf/1510.0478v1.pdf>

Glenn Fleishman, How Do Genome Sequencing Centers Store Such Huge Amounts of Data?

<https://www.technologyreview.com/s/542806/how-do-genome-sequencing-centers-store-such-huge-amounts-of-data/>

Petagene Company Website, <https://www.petagene.com>

Claire Giordano, Six Reasons to Add Object Storage to Your Genomics Lexicon,

<https://www.biosciencetechnology.com/article/2015/12/six-reasons-add-object-storage-your-genomics-lexicon>

Vivien Marx, Biology: The big challenges of big data, <https://www.nature.com/articles/498255a>

Nicola Davis, 'Angelina Jolie effect' boosted genetic testing rates, study suggests,

<https://www.theguardian.com/science/2016/dec/14/angelina-jolie-effect-boosted-genetic-testing-rates-study-finds-breast-ovarian-cancer>

---

Craig Knighton, 5 Opportunities and Obstacles in Genomics and Personalized Medicine,  
<https://mentormate.com/blog/optimizing-genomics-and-personalized-medicine/>

BCG Report, How Genomics and Genetics are Transforming the Biopharmaceutical Industry,  
<https://www.bcg.com/documents/file13745.pdf>

Genetic Alliance UK, Benefits and Risks of Genetic Testing, <http://www.geneticalliance.org.uk/information/services-and-testing/benefits-and-risks-of-genetic-testing/>

Centogene Website, Benefits of Genetic Testing,  
[https://www.centogene.com/fileadmin/pdf/Patients/Genetic\\_testing/Patient\\_information\\_Benefits\\_of\\_genetic\\_testing.pdf](https://www.centogene.com/fileadmin/pdf/Patients/Genetic_testing/Patient_information_Benefits_of_genetic_testing.pdf)

Monitor Deloitte, An Industry Study for the Office of Life Sciences,  
[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/464088/BIS-15-543-genomics-in-the-UK.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/464088/BIS-15-543-genomics-in-the-UK.pdf)

Frida Holme, Blockchain—The End of Privacy Concerns in Genomic Data Handling?,  
<http://www.frontlinegenomics.com/review/18569/blockchain-genomic-data/>

Canadian Genomics Cloud Whitepaper, <https://genomicscloud.ca>

San Diego Regional EDC, Cracking The Code: The Economic Impact of San Diego's Genomics Industry,  
[http://www.sandiegobusiness.org/sites/default/files/Cracking%20the%20Code%20-%20The%20Economic%20Impact%20of%20San%20Diego%27s%20Genomics%20Industry\\_0.pdf](http://www.sandiegobusiness.org/sites/default/files/Cracking%20the%20Code%20-%20The%20Economic%20Impact%20of%20San%20Diego%27s%20Genomics%20Industry_0.pdf)

Charlie Osborne, Dubai to DNA sequence its entire population, <https://www.zdnet.com/article/dubai-to-dna-sequence-its-entire-population/>

Geneformics Press Release, <https://www.businesswire.com/news/home/20170807005261/en/Geneformics-Announces-Scalable-Genomics-Data-Compression-Solution>

Genetics: From Genes to Genomes, Canadian Edition, McGraw Hill Ryerson, ISBN-10: 0070946698;

Think You Know Human Genome Sequencing Pros and Cons? Think again!,  
[http://www.jwms.reg4.k12.ct.us/UserFiles/Servers/Server\\_177389/File/Genome%20Sequencing%20Articles.pdf](http://www.jwms.reg4.k12.ct.us/UserFiles/Servers/Server_177389/File/Genome%20Sequencing%20Articles.pdf)

Markets and Markets Press Release, <https://www.marketsandmarkets.com/PressReleases/genomics.asp>

Hexa Research, Genomics Market Share, Size, Analysis, Growth, Trends and Forecasts to 2024,  
<https://pt.slideshare.net/marshwilliam98/genomics-market-research-report-global-industry-analysis-size-share-growth-trends-and-forecast-2020>

Jack Rudd, Managing the Genomics Data Deluge,  
<https://www.technologynetworks.com/informatics/articles/managing-the-genomics-data-deluge-288315>

---

HGST Use Case, Meeting the Storage Challenge of Exponential Data Growth in Genomics, [https://www.hgst.com/sites/default/files/resources/Meeting\\_the\\_Storage\\_Challenge\\_of\\_Exponential\\_Data\\_Growth\\_EN\\_US\\_1015\\_UC01.pdf](https://www.hgst.com/sites/default/files/resources/Meeting_the_Storage_Challenge_of_Exponential_Data_Growth_EN_US_1015_UC01.pdf)

Olivia Judson, Testing genes, solving little, <https://www.nytimes.com/2008/08/18/opinion/18iht-edjudson.1.15388463.html>,

Michelle Martin, Weighing pros, cons of genome sequencing, <https://www.osv.com/Magazines/TheCatholicAnswer/TCAFaith/Article/TabId/822/ArtMID/13670/ArticleID/10283/Weighing-pros-cons-of-genome-sequencing.aspx>

U.S. National Library of Medicine, What are the risks and limitations of genetic testing?, <https://ghr.nlm.nih.gov/primer/testing/risklimitations>

Christina Farr, Why Patients Are Getting Hit With Surprise Bills After Genetic Testing, <https://www.fastcompany.com/3059072/why-patients-are-getting-hit-with-surprise-bills-after-genetic-testing>

H. Gilbert Welch and Wylie Burke, Why whole-genome testing hurts more than it helps, <http://www.latimes.com/opinion/op-ed/la-oe-welch-problems-predictive-medicine-20150428-story.html>

Anna Nowogrodzki, Blockchains Won't Fix the Problem with Genomics, <https://medium.com/neodotlife/blockchains-and-genomics-32fc64fbb8f0>

Swiss Re, How Genetics Testing Will Impact Life Insurance, [http://institute.swissre.com/research/library/Genetics\\_Seeing\\_the\\_future.html](http://institute.swissre.com/research/library/Genetics_Seeing_the_future.html)

The Economist, Genetic testing threatens the insurance industry, <https://www.economist.com/finance-and-economics/2017/08/03/genetic-testing-threatens-the-insurance-industry>

Kate Rogers, Genome Sequencing: Who Gets to Use the Data?, <https://www.foxbusiness.com/features/genome-sequencing-who-gets-to-use-the-data>

Sabine VanderLinden, Genomics & Insurance: The Rising Tide of Genetic Data, <https://www.startupbootcamp.org/blog/2016/10/genomics-insurance-rising-tide-genetic-data/>

Warren S. Hersch, Genomics: upending the insurance world (for the better), <https://www.thinkadvisor.com/2016/05/12/genomics-upending-the-insurance-world-for-the-bett/>

Robert Klitzman, MD, Paul S. Appelbaum, MD, and Wendy Chung, MD PhD, Should Life Insurers Have Access to Genetic Test Results?, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4259574/>

Achim Regenauer, Chief Medical Officer, The Rising Tide of Genetic Data – New Challenges for Life & Health Insurance, [http://partnerre.com/opinions\\_research/the-rising-tide-of-genetic-data-new-challenges-for-life-health-insurance/](http://partnerre.com/opinions_research/the-rising-tide-of-genetic-data-new-challenges-for-life-health-insurance/)

Leslie P. Francis, Genomic knowledge sharing: A review of the ethical and legal issues, <https://www.sciencedirect.com/science/article/pii/S2212066114000283>



---

Michael Le Page, The ethics issue: Should we edit our children's genomes?,  
<https://www.newscientist.com/article/mg23531330-700-the-ethics-issue-should-we-edit-our-childrens-genomes/>

Norman A. Paradis, Your Genome May Have Already Been Hacked, <https://www.livescience.com/62442-your-genome-hacked.html>

Nucleotides and the Double Helix, [http://cyberbridge.mcb.harvard.edu/dna\\_1.html](http://cyberbridge.mcb.harvard.edu/dna_1.html)  
Wikipedia DNA Sequencer, [https://en.wikipedia.org/wiki/DNA\\_sequencer](https://en.wikipedia.org/wiki/DNA_sequencer)

Tibi Puiu, How big is a petabyte, exabyte or yottabyte? What's the biggest byte for that matter?,  
<https://www.zmescience.com/science/how-big-data-can-get/>

Wikipedia, Human Microbiome Project, [https://en.wikipedia.org/wiki/Human\\_Microbiome\\_Project](https://en.wikipedia.org/wiki/Human_Microbiome_Project)

Liren Huang Jan Krüger Alexander Sczyrba, Analyzing large scale genomic data on the cloud with Sparkhit,  
<https://academic.oup.com/bioinformatics/article/34/9/1457/4747885>

Arador, Comparing Bandwidth Costs, <https://arador.com/ridiculous-bandwidth-costs-amazon-google-microsoft/>

Lisa Zyga, New quantum repeater paves the way for long-distance big quantum data transmission,  
<https://phys.org/news/2018-02-quantum-paves-long-distance-big-transmission.html>

T1 Shopper, Online Calculator, <http://www.t1shopper.com/tools/calculate/downloadcalculator.php>

Frank A. Feltus, Joseph R. Breen, III, Juan Deng, Ryan S. Izard, Christopher A. Konger, Walter B. Ligon, III, Don Preuss, and Kuang-Ching Wang, The Widening Gulf between Genomics Data Generation and Consumption: A Practical Guide to Big Data Transfer Technology, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636112/>

Wikipedia, Sunway TaihuLight, [https://en.wikipedia.org/wiki/Sunway\\_TaihuLight](https://en.wikipedia.org/wiki/Sunway_TaihuLight)

Brian Westover, What Is Gig-Speed Internet?, <https://www.tomsguide.com/us/gig-speed-internet,review-5134.html>

Fiona Macdonald, Researchers Have Achieved Wireless Speeds of 1 Tb Per Second,  
<https://www.sciencelert.com/researchers-have-achieved-wireless-speeds-of-1-tb-per-second>

Timothy Prickett Morgan, Details Emerge On "Summit" Power Tesla AI Supercomputer,  
<https://www.nextplatform.com/2016/11/20/details-emerge-summit-power-tesla-ai-supercomputer/>

## Disclaimer

© 2018 The Hutch Report

The content of The Hutch Report and its web site [www.thehutchreport.com](http://www.thehutchreport.com), is provided for information purposes only. No claim is made as to the accuracy or authenticity of the content found in this report or on our website.

The Hutch Report is not a Registered Investment Advisor, Broker/Dealer, Financial Analyst, Financial Bank, Securities Broker or Financial Planner. The Information in this report is provided for information purposes only. The Information is not intended to be and does not constitute financial advice or any other advice, is general in nature and not specific to you. You should seek the advice of a qualified and registered securities professional and undertake your own due diligence before making any investment. None of the information on this report is intended as investment advice, as an offer or solicitation of an offer to buy or sell, or as a recommendation, endorsement, or sponsorship of any security, Company, or fund. The Hutch Report is not responsible for any investment decision made by you. You are responsible for your own investment research and investment decisions.

The Hutch Report website and reports/newsletters do not accept any liability to any person or organisation for the information or advice (or the use of such information or advice) which is provided on its web site or reports/newsletters or incorporated into it by reference. The information on the reports/newsletters and the website is provided on the basis that all persons accessing the site undertake responsibility for assessing the relevance and accuracy of its content.



[www.thehutchreport.com](http://www.thehutchreport.com)