Deep Science Publishing
https://doi.org/10.70593/978-93-49307-76-6_2

DeepScience
Open Access Books

Chapter 2: The role of big data in genetic testing: Infrastructure, data lakes, and ethical management of genomic information

2.1 Introduction

The massive funding that has been invested in genomics has made it a field of data science fundamentally allied with computer science. A consequence has been the development of an infrastructure of open-access public repositories such as the Sequence Read Archive, the European Nucleotide Archive, or the DNA Data Bank of Japan. Such repositories are essential for the advance of the field by efficiently storing and sharing genomic information while ensuring privacy protection and data confidentiality. However, it is also necessary to anticipate potential ethical risks according to the newest multidisciplinary paradigm in genomics. Firstly, security hazards are expected, such as the kind of attack articulated distributed denial of service known as ‘doxing’. Other genetic information management efforts must be directed toward the strict regularization of the treatment and storage of genetic data and to preventive endpoint measures. In this last respect, it is of utmost importance that genetic data is never directly part of the final outputs without anonymization, but rather that it is accessed through other intermediate automation or analysis. Nonetheless, the desirability of these data will certainly attract the development of an artificial intelligence irretrievable watermarking layer of data storage peripherals, with direct blocking mechanisms installed in special hardware components of networks or climax computing servers.

Each cohort member is genotyped at a set of Single Nucleotide Polymorphisms (SNPs) present in a microarray-based chip. Around one million SNPs per individual are genotyped giving rise to the impressive volume of data to be analyzed. These SNPs are spread all over the genome, so in turn, genetic data management involves linking these genetic variants to biological features, such as genes or pathways. In addition,

genotyping several individuals at a time leads to hard computational effort in terms of data management and manipulation. Therefore, the analysis of such data is running on infrastructures hosted by High Performance Computing (HPC) centers. Since most researchers have no access to these Computer Facilities, the need for user friendly gateways is high on demand. Most data and results are kept in public repositories. Data management issues range from the internationally accepted form of sharing GWAS datasets to the design of better infrastructures or data lakes, able to hold and relate the large number of entities in use. A wide range of services are offered to the academic community: analysis-ready datasets, tools for data manipulation, lists of known associations or the post of unpublished research results. A specific kind of data that is being stored and shared refers to genetic codification itself. Care must be taken to ensure privacy protection and data confidentiality, which makes this kind of data less openly available.

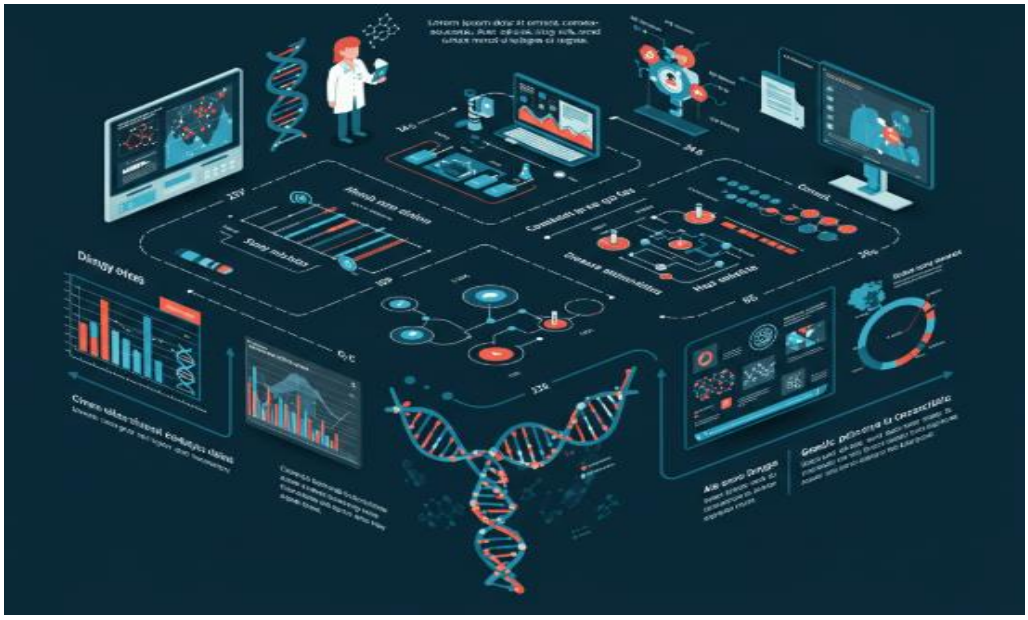

Fig 2.1: Genetic big data

2.1.1. Background and Significance

New technologies have enabled genetic testing at the population level, fostering programs that strongly rely on the incorporation and application of genomic information. Thus, large amounts of genetic data, mostly in the form of genotypes have to be analyzed. This data avalanche poses new challenges in terms of scalability and, especially, in terms of data handling, storage, dissemination and privacy management. Consequently, current genetic studies are experiencing a radical shift in the

methodologies in use, being computational facilities the bottleneck of this change. It is critical that the academic and health community consider this challenge and be prepared for it in advance, or this enormous amount of valuable information will be useless or, at the very least, its potential will be greatly limited. Large cohorts are being genotyped and the data is explored with the ultimate goal of understanding the genetic bases of complex traits. A study of this kind, known as Genome-Wide Association Study (GWAS), analyses the genomes of a large number of individuals searching for statistical associations between genotyped variants and the phenotype being studied. This strategy has superseded the Candidate Gene approach used so far, mainly because not many genes were found to be significantly associated with complex traits. In any case, GWAS remains an intense area of research, considering that it was first proposed less than a decade ago. With the aim of sharing costs and enhancing statistical power, several consortiums have been launched.

2.2. Understanding Big Data

To begin with one should go through Big Data, as the infrastructural and legal aspects of Huge Data are required for the efficient large-scale management of genomes. Thereafter, one might proceed to describe the architecture of a "Data Lake" and the use of neural networks while this infrastructure is engaged in managing genetic data. As a consequence, a brief commentary will be made on the privacy issues of genetic information. Introductory remarks about promising tools for managing big data to the benefit of genomic analysis and about current projects working in both fields are inline. As part of the introduction, it's worth noting that public DNA databases have increased genome-wide association study power and that ethicists are adapting as the field evolves. As methods and technology evolve, it is increasingly important to have those best equipped to use it engaging with data producers to assist in planning and extraction. Emerging from the field early on, ethicists were able to aid in the compiling of policy and legislation before the data services were established, such as the Hippocratic Oath. Scalars and pharmacogenomics are just two examples of the ways in which genetics can digitally inform healthcare. Public repositories of genetic and environmental influences on health for the wider research and policy communities are essential for this advance. Such an endeavour requires efficiently storing, retrieving, and sharing data while ensuring privacy protection and data confidentiality. Therefore, considering genetic information and biometric markers; including behavior, health, education, finances, geo-tracking, and other data, also cloud dedicated communication software targeted on hacking prevention. Ethical risks anticipating potential security hazards deriving from the storage and sharing of genomic data; risks that may involve various actors. However, those associated with genomic data are expected to be lower than in other sources of biomedical and health-related information.

2.2.1. Definition of Big Data

In biology, the emerging field of big data involves the application of technologies that manage, store, and analyze the extensive data sets produced by the process of genetic testing. This initiative seeks to understand how a person's genetics, environment, and lifestyle can make them more or less susceptible to any given ailment. Information technology, and big data in particular, will be central to this initiative. In the realm of genetic testing itself, perhaps the biggest obstacle to the widespread implementation of genetic information has been the difficulty of efficiently storing, retrieving, and analyzing huge data sets. As technology increasingly evolves towards enhancing the granularity of this genetic information, the need for scalable and efficient information technology will only increase. Let's take a closer look at the broader range of technological developments within the big data domain impacting the field of genetic testing (Komaragiri, 2022; Chakilam, 2022; Challa, 2023).

Turning genetic information into actionable knowledge requires the broad development of these cyber infrastructures. Big data in the realm of genetic information has recently been characterized as an 'infrastructure problem'. Within the domain of genetic testing, the remote movement and storage of such individual genetic data has been, and in some cases continues to be, so overly burdensome as to render the operation of genetic testing largely impractical. In order for the future of genetic testing to live up to its promise, this "information superhighway" of genetic information needs to be created. The growing collaboration between information technologists and genetics specialists has begun to effect significant progress towards this end, though the big data-associated difficulties remain numerous. It has begun to significantly aid the process of data warehousing, efficient data transfer, data compression, and bioinformatics. An increasingly exciting format of this information technology-gene force is the creation of data depositories known as data lakes that facilitate the storage of all forms of genetic data in a flexible format that allows for seamless analysis. These public repositories are being seen as the key to allowing for the broad advancement of genetic medicine. On the other hand there is an arising debate on the ethical management of this immense flow of genomic information.

2.2.2. Characteristics of Big Data

Technological advances have enabled the successful detection of genetic variants causing hereditary disorders but the sheer volume and complexity of the generated genetic data create the need for specialized workflows and information systems. In this sense, the role of Big Data when dealing with this issue cannot be understated. However, legal and economic issues arise. Firstly, the infrastructure must be established in order to accumulate the sequenced genetic information. It is either the hospital or the primary

care centers that must take care of this, and many will not have the possibility to do so. It remains to be seen whether the healthcare sample collections are privatized, and how much sharing of that information takes place between public and private actors. Caps will be set for the fees that can be charged. This will concentrate the information in private biobanks (or encourage collection donation). Fees related to donors' data disclosure will be regulated as well. Despite the ambition to serve as a blueprint for other member states, the draft legislation may stigmatize the kind of data-mining practices it aims to facilitate. Secondly, the creation of the data pipelines in such a way that it matches the growing understanding of human biology, meaning that datasets originating from different technologies and subject populations would have to be combined and analyzed in concert. Data lakes will have to be established for this purpose; the generation of such metadata-enriched data lakes will likely remain confined to the big pharma companies. Infrastructure on this scale is unlikely to be available in secondary care, and specialty care will be restricted to only a few areas. It is possible that for some rare disorders it will cost less to sequence the whole genome of every newborn than to continue paying for the diagnostics. This one-off effort is likely to lead to vast biobanking of genetic data (Malempati, 2022; Nuka et al., 2023).

2.2.3. Sources of Big Data in Healthcare

In the healthcare industry, various sources for big data include hospital records, medical records of patients, results of medical examinations, and devices that are part of the internet of things. Biomedical research also generates a significant portion of big data that is relevant to public healthcare. Patients have also begun generating a relatively large amount of data about their health, either through wearable devices or mobile applications. Of all the contributions to big data in healthcare, genomes are generating potentially the biggest interest. It is expected, and actively encouraged, that people will have their genomes sequenced, generating huge amounts of new information. This information can help to diagnose and select treatments for diseases; to inform what genes may be passed onto offspring; or even to improve physical condition. It is also an interesting basis for historical and genealogical studies. While most big data in healthcare is created by machines with digital records, patient-generated data is usually analogue. However, technologies allow this data to be digitized and fed into algorithms. All of these data sources have repercussions in public healthcare that, properly dealt with, provide all sorts of opportunities. But at the same time, great challenges are required. Furthermore, the openness with which biomedical industries treat the aforementioned data sources has caused controversies that show the growing concern about the ethical treatment of patient data. The management of this data involves a huge set of recommendations that are outlined here.

2.3. Genetic Testing Overview

Since the early 1990s, genomics has grown in importance for the purpose of understanding and treating health conditions. Genome sequencing technological advances have precipitated a reduction in the cost of sequencing and have led to the commercialization of genetic tests offered by an increasing number of companies. Genomics is becoming sophisticated and has the potential to provide predictive health screening tools capable of identifying a predisposition to specific diseases later in life. Having your genome analysed might soon be as routine as having your blood pressure taken.

Genomic data is sensitive and can reveal information about a person's health, susceptibility to disease, ethnic background, paternity, and relationship to other individuals. There are vital ethical and legal questions emerging from the storing, sharing, and retrieval of genomic data. Genomics raises peculiar questions since the relevance and implications of part of the information that can be gained from the analysis of genomic data might be expected to be discovered in the future. There are a number of existing ethical guidelines and legislation with respect to the use of genomic data in clinical practice, genetic counselling, or research studies with human subjects. Nonetheless, there are significant gaps in the existing policy regulating the cross-sector use of genomic data between the public health system and the private sector as well as between the biomedical companies and third sector organizations or sectors outside the health system.

However, in the nascent industry, the data storage resources required for the broad-scale adaptation of these approaches, including tissue-specific quantitative trait loci databases, are not readily available. Systematic sharing of genotypes files within the research community has already begun and it has proven to be essential in different research areas. Such public repositories are essential for the advancement of the field as they allow researchers to efficiently store and share the genomic information while ensuring the application of a set of practices for privacy protection and data confidentiality. Moreover, research projects that need to access the same genomic information can benefit from a coordinated strategy and standardization.

2.3.1. Types of Genetic Tests

A genetic test is a kind of medical test that looks for changes in a person's genes, or searches for possible health problems caused by these gene changes. First of all, many people take genetic testing to find out whether they have a genetic condition or to see if they have a higher risk. While the test can provide useful information, those results are not always clear or simple. It may be burdensome to make decisions. Genetic testing has

some limitations. For example, it cannot answer all the questions regarding certain genetic conditions. In some cases, a positive result does not say what to expect fully. At the same time, genetic testing has some risks. The test can prove uncertainty about developing a disease. In rare cases, someone might receive an incorrect result.

There are three major types of genetic tests. The first type looks for some kinds of changes in a person's genes, chromosomes, or proteins. The second sort of tests can give an idea of a person's chance based on a family health history. For someone who hasn't had a particular disease before, this can be more informative. The last category is faced with a certain disease or condition. This type of test may use different methods and look for different types of changes, based on this, there are several methods to distinguish them. For instance, methods for genetic tests include tissue sampling, blood draws, cheek swabs, and others. Moreover, the DNA from the sample is assessed using a test to seek significance. Techniques for taking genetic tests involve linkage testing, molecular testing, and so on. Finally, there are several key aspects that need more attention before and after tests. It can be the case that some questions should be asked. Therefore, readiness is key to becoming prepared for a check. On the other hand, a genetic counselor sometimes has to be met. Only when a test result is back do the results need to be understood.

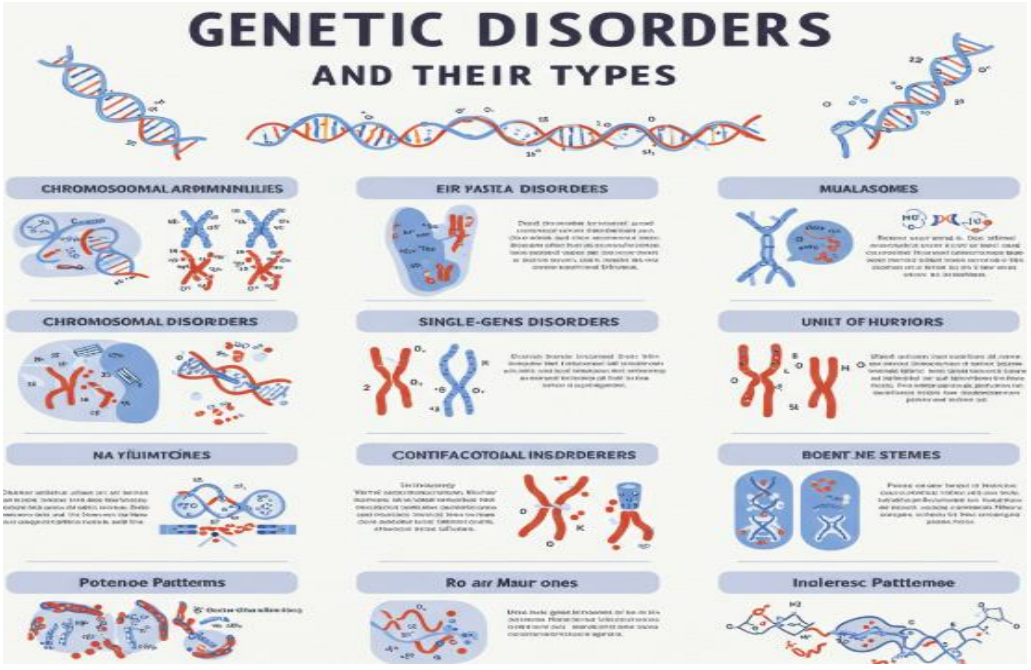

Fig 2.2: Genetic Disorder and Types

2.3.2. Applications of Genetic Testing

Genetic testing entails the analysis of DNA to identify variants associated with disease. Given that genetic information is passed from parents to offspring, families could benefit from whole genome sequencing (WGS) to share with their relatives, but do not have the means to store and interpret large volumes of data whilst maintaining privacy. Information in WGS is abundant: a typical dataset adds up to 50 gigabytes, which needs powerful infrastructure to store and share conveniently. Recording every individual's genetic information on their electronic health record (EHR) across a national health service would require vast amounts of storage and computational power. At present, it is not possible to handle this data using any traditional storage means available to healthcare providers. However, there is an attempt to visualize a design using sophisticated data lakes and APIs that could safely store and interpret large volumes of genomic data and EHR interactively. Genetic data are a subset of genomic information, comprising a person's entire set of DNA. Genomic data is abundant, consisting of up to 6bn base pairs. Further to this, genomic variation across individuals is vast, hence in comparison to other data types, genetic information requires considerably more storage to precisely represent the data. There are 4-12m genetic variants per individual, quickening the vast storage required. Genomic information is encoded in binary and stored in VCF files: the description of that variation in files is verbose and the data itself exhibits poor compression. Combined, this means that a single VCF file takes up a substantial amount of disk space. Given that the NHS requires four truthful storable representations of all genetic information over time, it could be expected that genomic storage will become a significant issue in the near future. Broad infrastructure would be needed to allow individuals to engage with their genomic information effectively, keeping privacy considerations front of mind. Bioinformatic algorithms could be developed to interpret genomic data, outputting the interpretation in an accessible format. Genomic information and its interpretation are shared between a clinical genome (EHR) and patient. Interested parties would require a secure, private means to store large volumes of genomic information for an extended time, as well as powerful computational resources to analyze it. This focuses on these concerns, setting out a vision for healthcare-linked genomic data lakes with ready-built algorithm APIs.

2.3.3. Advancements in Genetic Testing Technologies

This chapter starts off by outlining the big data investments being made over the month with $4 million in funding from the government budget in the development of new software to be built on existing infrastructure and hailed as a major game changer in the efforts to build a national cancer database. The second part of the chapter sets out key requirements for the project as a bid is due for pre-qualification for the second round of

ICT funding with a state health department as a clinician data provider. The remaining part of chapter 3 goes into tenants around retaining the connected electronic health record and talks about underground data lakes.

Big data investments have certainly been in the public eye over the past months. Some $4 million in funding, allocated from the previous year's government budget, has been identified in growth initiatives. Some of these include investment in the development of new software, allowing for more informative data to be gleaned from the existing infrastructure. This software will be used to produce, for the first time, hospital-level cancer-related death rates. It has been touted as a major game changer in the establishment of a national cancer database.

Since the funding announcement and as the system purchaser, the government has been keen to better understand what can realistically be delivered. In a call today it was asked how the system graduated from the proof of concept to a national database of cancer related deaths in all private and public hospitals. The first prerequisite is amendments to the framework agreement so that the system can be rolled out to all 51 public hospitals in the state. The issue at present, is that the system is not suitable for use in a live environment. This means that hospitals won't be able to use it until problems like accessing patient-identifiable data, privacy concerns and hospital firewall issues are rectified. The immediate commitment from all parties is for the proof of concept to be finalised. Responsibility for building the software will be retained by the current developers, with the steering committee to facilitate this.

2.4. Infrastructure for Big Data in Genetic Testing

How Can We Keep Big Data's Promises? Looking at the Infrastructure is a good start in exploring the topic. Biomedical science has evolved drastically in the last couple of years, providing frameworks easy enough to extract and analyze. Thanks to the genomics revolution, there are cheaper DNA sequencing devices able to read and obtain large DNA amounts. However, protecting confidentiality is extremely complex. Interestingly, in the US, people own their DNA and thereby their sequenced information. The arrival of DIY testing kits is also questioned due to genetic discrimination and privacy concernment. With massive DNA data coming on a big data platform, the ethical storage, processing, and sharing solutions of genetic information becomes fundamental.

5. Biomedical Science Goes Big? Better Build Data Lakes: Big Data Technologies and Practices Arising in the Biomedical Sciences is a suggestion and Safety in the Value Chain of Big Data is a Concern to Discuss. The enormous capacity to store genomic data will soon exceed the state of the art. Public data depositories have thrived, delivering extensively processed and cleanly structured data. Traditional relational data systems or

ad-hoc genomic stuff have proven efficient limitations. On the other hand, genomic data collect fine-grained data types in the form of sequences and time points, creating data silos problematic to render into a common flow. As genome Holland's revealed, genetic information can't be handled without attention to the ethical concerns it arouses, whose paramount concern is data safety. It consents to follow the authentic genotype providers up until they achieve integrated and analyzed measures.

2.4.1. Data Storage Solutions

There are three public repositories widely used for the storage of genome sequences: the Sequence Read Archive (SRA), the European Nucleotide Archive (ENA) and the DNA DataBank of Japan (DDBJ). SRA has been in epidemiology for some years, but these repositories have not been in operation for a long time.

If we want to properly manage genomic data it is of utmost importance to care not only for data and filename structures, but also for metadata documentation. Clinicians are currently reluctant to the idea of data sharing due to ethical reasons and that has made most relevant ethical guidelines for the sharing of genomic data to put an emphasis exclusively on patient consent. But there is also a pool of other stakeholders involved in the achievement and interpretation of genomic data that should be taken into consideration: hospital administrations as they own most of the material resources, bioinformaticians as the profiles responsible for data analysis and, most important, the professionals part of the empowerment-individual.

Some have proposed the implementation of secure distributed big data systems called data lakes to properly manage the accumulation for the analysis of large volumes of genomic information. Ten years from the Human Genome Project every genomic laboratory has an assortment of sequencing and genotyping platforms. All these public repositories require samples to be sent to third-party core sequencing facilities, which in the vast majority of cases are not hospital-owned. This information lies out of the legal reach of the health regional systems.

2.4.2. Computational Infrastructure

Genomic medicine requires high-performance computing networks. The generation and analysis of genetic data, in particular clinical sequencing, is of extremely high computational demand. Hence, the recent interest of the genomics community in network optimisation, either for cost effectiveness or performance speed-up. Low latency of network transmission has a direct impact in folder transfer when analysing huge amounts of the data.

As a consequence, a lot of the effort has been turned to the creation of data lakes, or repositories of genomic data. There are very ambitious projects for setting up huge data lakes of genomic listings, that constitute an opportunity to make significant progress in the comprehension of the secret of life. At the same time, privacy risks derived from the storage and sharing of genomic data make some scientists feel uncomfortable. About 1% of the human population supposedly has a relative whose genetic information is available in the GEDmatch database, consisting of a higher proportion of individuals of European ancestry. demonstrates that the availability of rare genetic diseases can approximately be derived for about 3% of the human population (1 in 33 subjects), exceeding 50% for larger groups. Closer relatives are more likely to be recognised because they share a larger number of segments. On average, 10% of the first cousins were inferred as such by the other participants, but short genetic segments were still exchanged for 40% of all pairs. All in all, five million more relatives (10% on top of the 50 million individuals already in the comparison) could be found on this database. The vast majority of these new inferences would be valid because the FNR at that kinship level was only 5.4%. These repositories are of key importance to advance in the field as they allow for efficient storage and sharing of genomic information. It is hence crucial to ensure these repositories are based on best practices that guarantee data confidentiality and privacy protection. Public information on ongoing initiatives that act as data lakes of genomic information should clearly state the measurements that are being undertaken to protect individuals' data and information on the kind of data that is stored and shared. In such a context, a set of best practices on how to set up such data lakes might help and foster the creation of repositories, adherent to strong privacy and data confidentiality principles.

2.4.3. Integration of Data Sources

Precisely, large datasets are needed in this context, and great efforts are being made to mine them effectively and a legal infrastructure can support data flows and analytics. In recent years a proliferation of consortia or public repositories has been witnessed. They share genomic data from large cohorts of individuals on a free and open access basis. These public repositories are essential for the advance of the field allowing research groups to focus on the study of the data rather than in the generation of genomic information while ensuring that genomic information is efficiently stored and shared along with the granting of privacy protection and data confidentiality. However, as the DNA of more patients is sequenced and as genomics takes a more prominent role in clinical decision-making, better management is needed in order to safeguard and explore the information contained in the genome. There is a gap in the healthcare sector because of the lack of data policies and practices around the use of latent genomic data. There is a debate spanning medical ethicists, doctors, computer scientists, and the industry which

is accentuated in light of the recent completion of the genome and the ongoing development of genome-wide tests.

There are also concerns about the potential misuse of genomic information. Although security hazards associated with genomic data must not be underestimated, they appear lower than those of other popular sources of biomedical and health-related information. In many cases, data is stored in a de-identified form, a necessary requirement for ethical and legal reasons but also it hinders straightforward exploits. In other occasions DNA data is complemented with 'low-dimensional' phenotypic information, such as single SNP-trait or markup. This fact makes the DNA information itself less attractive for unethical aims. What's more, on average, biometric data breaches have no consistent effect on individuals, especially when it refers to the sequencing of latent identical genomic information. Finally, genomic data is still less tempting than other datasets to attract a substantial amount of interest apart from research and clinical scenarios.

2.5. Data Lakes in Genomic Research

There is growing awareness that the commercialization of genome sequencing is likely to create a novel form of big data. While individual genome sequencing files are comparably small in size (of the order of a few gigabytes each, often less), it is metadata describing the genetic information they contain which is truly commodified. As a consequence, the rapid commercialization of genome sequencing is giving rise to an unprecedented expansion of the genomics big data sector – a recent report projected a market increase from $1.35 billion in 2016 to $8.61 billion by 2022, representing a compound annual growth rate (CAGR) of almost 30%. Much of this market is characterized by business-to-business transactions wherein genomic information is sold and resold with the purpose of licensing or selling access to biobanks, analysis pipelines and data lake systems. At present, bioinformatics conglomerates are working to establish a new, global infrastructure that directly rivals well-known public repositories. At the time of writing, the biggest is based on various agricultural datasets (several quadrillion petabytes worth) and it is run by a consortium of companies. These bioinformatics conglomerates do store a minority of human-derived sequencing files, but the vast majority of interest lies in the trading of value-added genomic information from proprietary biobanks – the reason they have been hard at work aggressively lobbying governments with the aim of loosening biobank regulations.

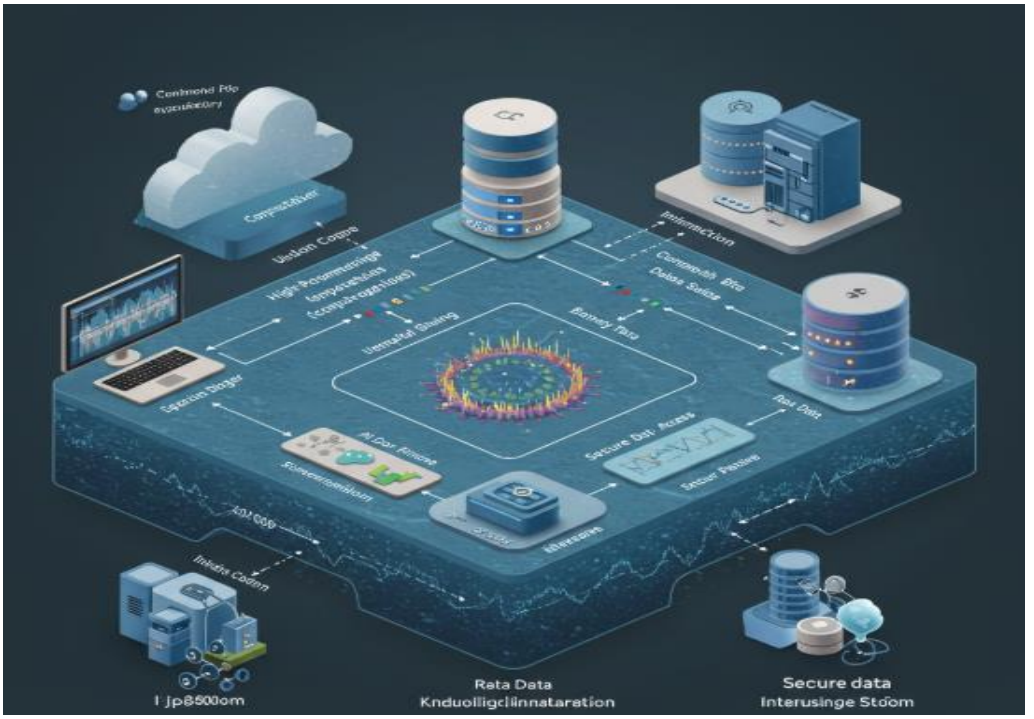

Fig : Data science infrastructure for genomics data

2.5.1. Concept of Data Lakes

In this context, the concept of data lakes is important when we think about storing and integrating collections composed of big data. Data lakes could be thought of as heavily indexed data warehouse services that keep different collections of large-scale data assets, such as genomic and health data, alongside a range of more classical data storage options. There are data lakes where the emphasis is on how to provide them as a service, or as a federation of services. From a genomics point of view, public data lakes will be of interest, such as. These public repositories are necessary for the field as they make it possible to store and share genomic information efficiently, while carefully considering patient specifications in order to maintain privacy protection and data confidentiality. On the other hand, there can be in-house data lakes, often built around particular hardware and software stacks, and often with additional access requirements, agreements and/or costs. For example, cloud- or on-premises-based platforms having a collection of (de-identified or otherwise-defined) genomic and related health data assets, available for specific researchers or clinicians.

As technical infrastructures for healthcare genomics advance, it is useful to consider the subject of how most big data resources of genomics are dealing with. For instance, it is

well-known that a popular use of cloud resources is to provide auto-scaled web applications that can deal with wide and unpredictable workloads. Data analytics platforms are the most important elements when dealing with computational genomics and can either be specific, such as biomedical image or next-generation sequence processing, or domain-independent, such as data lakes, federated platforms, or HPC setups. The cost-efficiency of a solution with respect to the specific workload, mainly in terms of response time and ability to avoid idleness, is of paramount importance. Additionally, it is crucial to ensure compliance with health data regulations, safeguarding such data both during computation and during storage, longer or not.

2.5.2. Building a Data Lake for Genomic Data

Big Data infrastructures typical data storage facilities in fields like genomic research were explored. A technical outline was provided about the building of a data lake for genomic data that should be scalable, efficient, fault-tolerant, secure and easily accessible by multiple workflows. The emerging need for ethical guidelines in the big data handling of genomic information was also outlined.

In the last decade the advent of next-generation sequencing (NGS) technologies has revolutionized genomic research and healthcare. The steadily plummeting costs and the increasing throughput of NGS have led to vast amounts of DNA data that require ever-sophisticated computational post-processing to allow researchers and healthcare professionals to extract useful insights. Genomic data are of diverse types and sizes, with the raw sequencing data being larger and the most time-consuming to analyze computationally speaking. Like many other scientific data, genomic data are usually stored and shared among researchers using a set of public repositories. These public repositories have become an essential infrastructure for the advance of the field as well as to share and store genomic information in an efficient manner while guaranteeing protection of privacy and confidentiality. These computational post-processing steps have brought forth a new field at the intersection between biology and computer science, bioinformatics. NGS comes with its own set of challenges to bioinformatics due to the unevenness and complexity of the sequencing process. Adding a layer of complexity is the fact that for many applications, such as the routine processing of patient samples in hospitals, NGS results need to be obtained and reported as soon as possible. To address these challenges, many tools have been developed for the analysis of a particular type or feature of the genomic data.

2.5.3. Advantages of Using Data Lakes

Instead of storing data in a hierarchical way with relational, columnar or even hierarchical databases stored in so, in the so-called three Vs approach, data can be stored unstructured in flat files. There are four key advantages of this approach: • Data Lakes store data for several years in a secured way. They allow data to be stored in a raw format as it arrives, and interpretation and analysis can be put off until needed. Hence it is not necessary to know beforehand what to do with the data. • This infra-structure, based on flat files, has no single point of failure. If a data-lake goes down, the data is safe, and there is still time to set up another data-lake. Therefore, the effort of setting up data-lakes in the first place is securely invested. • Data-lakes are not just set up for the raw data to comply with data protection laws. Flat files in a data lake are always encrypted. In contrast to the most widely used alternative to store big data, data-lakes are on machines that do not have internet access. Therefore, the risk of data being transferred accidentally, or in the form of metadata, to third parties is much lower. • Data-lakes save data in a way that is much faster to access than a data-base. Tools are provided to access data as well as a user-friendly web-console.

2.6. Conclusion

Article is devoted to the current state of big data in the field of genomic research and commercial genetic testing. This paper offers an outline for the ethical management of genomic information and possible policy planning. Emerging public, private and mixed infrastructures of organizational systems, cloud computing, and Internet of Things (IoT) networks are explored. The advent of genetic testing has brought with it the unintentional generation of large scale genomic information. Commercial companies producing genetic tests have therefore had to adapt to the progress in infrastructure, communication and in the manipulability of large sets of disparate data. In addition to commonly considered aspects such as the biochip technology, robust algorithms and secure hosting services, handling of big genomic data has also involved networking scientists, storage scientists, e-infrastructure developers, and regulatory employees. To date, a triad of organisational setups or infrastructures can be distinguished: centralized data lakes, partitioned but interoperable public data resources, private resources with different proprietary security levels.

Direct-to-consumer genetic testing has established a significant business for companies providing it and has simultaneously presented these companies with questions about the use and storage of the generated genomic information. It is currently urged that the produced data be treated in a manner combining both patriotism and precaution. As such, vast responsibilities regarding the surveillance of biobanks and communicated privacy policies of commercial partners have arisen. By accepting a DNA-test, the consumer or

patient is allowing the genetic testing company to create vast amounts of genomic and genetic information about him or her. Traditionally, this information is condensed into a profile, which is made available for the customer, recommending more or less serious lifestyle advice. Nevertheless, additional measures can be made to the produced genomic information, whether before or after profiling.

2.6.1. Future Trends

Genetic testing is advancing rapidly towards the incorporation of large-scale sequence information, producing an explosion of Big Data. This situation raises new challenges in terms of the necessary computation infrastructure for data storage and computational analysis. Platform media schemes running on the open-source Hadoop framework that addresses the management of this vast amount of data are discussed. The suitability of the data lakes concept promoted by big industry regarding the facility and standardization of protocols for data transfer between clinics and analysis platforms as well as potentialities for a shopping-style approach to genetic testing are also considered. Finally, a strategy based on pseudonymization techniques in the elaboration of reports to patients that permit the management of massive genetic information while preserving the right to privacy is outlined.

Though joining the clinical setting has been more challenging, genetic testing is currently becoming part of the routine assortment of procedures performed by human physicians and veterinarians alike. The spread of new-generation techniques and sequencing devices has been accompanied by a progressive reduction in its cost. This has already ignited a revolution in the field of animal breeding, where the amount of information that can be provided for the selection of individuals has become enormous. Evolution in the clinical area has so far been slower due to its more demanding quality requirements and ethical concerns. However, change is also happening, and the incorporation of large-scale data analysis through sequencing technologies is now a reality in a small but ever larger number of clinics. In fact, the rising of commercial enterprises that are devoted to promising the complete exome or genome sequence of the individual represents a turning point subject to cross-disciplinary implications .

References

Komaragiri, V. B. (2022). AI-Driven Maintenance Algorithms For Intelligent Network Systems: Leveraging Neural Networks To Predict And Optimize Performance In Dynamic Environments. Migration Letters, 19, 1949-1964.

Chakilam, C. (2022). Generative AI-Driven Frameworks for Streamlining Patient Education and Treatment Logistics in Complex Healthcare Ecosystems. Kurdish Studies. Green Publication. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3719.

Malempati, M. (2022). AI Neural Network Architectures For Personalized Payment Systems: Exploring Machine Learning's Role In Real-Time Consumer Insights. Migration Letters, 19(S8), 1934-1948.

Challa, K. (2023). Transforming Travel Benefits through Generative AI: A Machine Learning Perspective on Enhancing Personalized Consumer Experiences. Educational Administration: Theory and Practice. Green Publication. https://doi. org/10.53555/kuey. v29i4, 9241.

Nuka, S. T. (2023). Generative AI for Procedural Efficiency in Interventional Radiology and Vascular Access: Automating Diagnostics and Enhancing Treatment Planning. Journal for ReAttach Therapy and Developmental Diversities. Green Publication. https://doi. org/10.53555/jrtdd. v6i10s (2), 3449.