Data Science in Genomics

Ajeigbe Oluwafemi

Washington State University

A B S T R A C T

Data science has allowed many industries, especially the health care industry extract practical insights from large-scale data [9]. The birth of an interdisciplinary field of study that enabled researchers to use machine learning and computational and statistical models to analyze available information such as mutations hidden in gene sequences have helped the health industry in many ways is referred to as genomic data science. New sequencing technologies like whole-genome sequencing and whole-exome sequencing targeted sequencing techniques are currently used in biomedical studies and medical practices to detect disease and drugs associated with genetic variations in advanced precision medicine [2,3]. One of the common challenges discovered while reading past works on genomic sequencing was the need to store, manage and analyze this sequenced data effectively [8, 10]. This review paper will survey how genomics data differ from data in other data-driven disciplines and its prominence amongst biological sciences.

Keywords: genomic analysis, genomic data science, big data

I N T R O D U C T I O N

Data science fuses various fields, including statistics, machine learning, programming, exploratory data analysis, and a couple of other techniques, to extract valuable insight from large-scale data [8]. Over the years, many disciplines have realized how vital data science is because data generated from these disciplines are usually not valuable and interpretable [8]. Data science was then re-vamped as it uses other disciplines to provide techniques for analyzing large-scale data while providing insights, patterns, & trends. Industries have used data science due to its commercial utility, and over time, discoveries, decisions, and inventions have been made using results from these data science processes [9, 10]. An article published by Dhar (2013) talked about how big data helped data scientists develop algorithms machines use to ask and validate interesting questions humans might not consider. Machine learning and predictive modeling have been instrumental in the healthcare industry as it has helped identify many causes of diseases and cures [9,10]. Researchers refer to the field of study in which genomic data science uses computational and statistical techniques to learn important hidden messages in genome sequences, as genomic data science in which studies are being done on genomes and how data science is used to understand sequences in genes [9,10].

With advances in next-generation sequencing technologies, the amount of sequence data generated is tremendous [8-10] Various research interest in the role of data science in genomics in the past decade has been fueled by the need to manage, query, and analyze the large volume of data being

generated [9]. The human DNA comprises about 3 billion pairs, representing 100GB of data [10]. It is, therefore, necessary to store, query, and analyze these big data from genomics, as reports have shown how vital the discoveries made from big data are. In the last decade, health and medical sciences are now in the post-genomic season, resulting from the exploration of new genomics sciences [9-10]. The new genomic sciences are made of technology encompassing complete genomic sequencing that can analyze sequences of tens of thousands of genomes. These new genomic technologies have rapidly increased generated sequences, creating significant resistance to the computing infrastructure and software algorithms used for genomics data analysis. Reports from the NCBI showed that genomic records in the GenBank database have exponentially increased over the last decade, doubling every 18 months. The figure below shows the growth of GenBank between 1989 and 2019. [10]

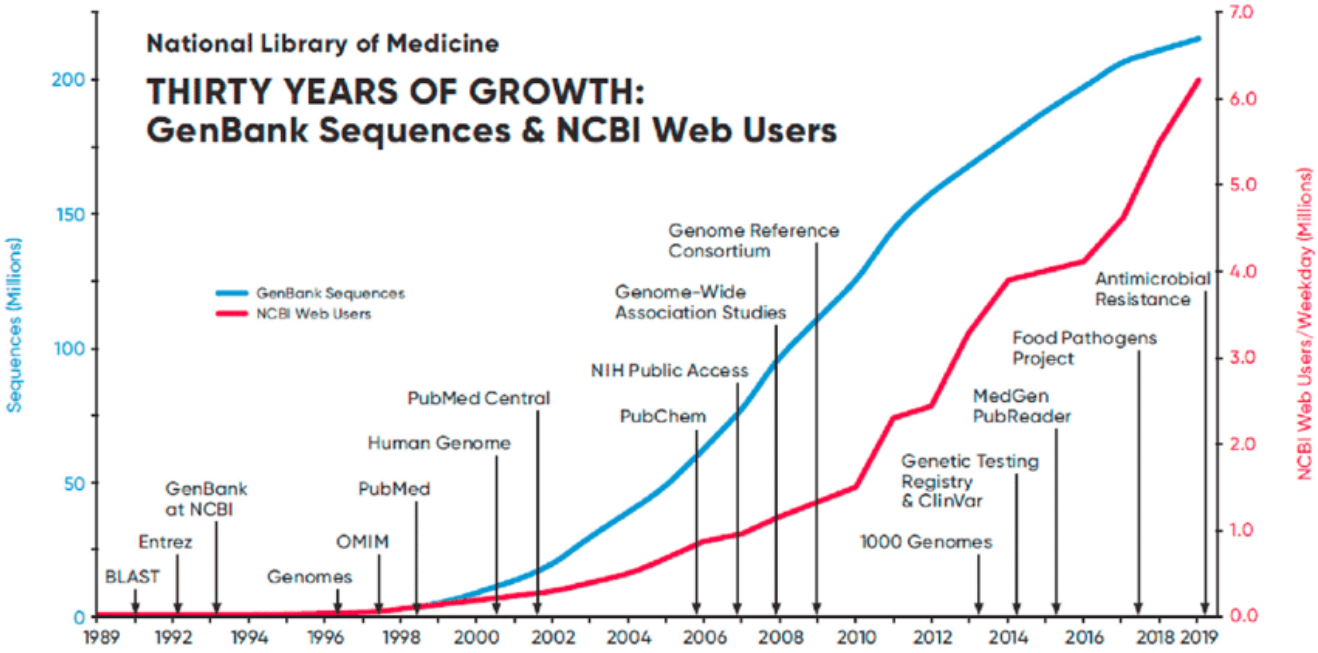

Figure 1: Thirty years of Growth: GenBank Sequences & NCBI Web Users
(Source: DHHSNIH (https://www.nlm.nih.gov/about/2021CJ_NLM.pdf).)

The big genomic data ecosystem model (Plamenka B., 2017) in Fig. 1 is a diagram that shows the various sources of genomic data. One of the highlights of the study by Plamenka examined the need to address the new challenge created by the rate at which the genomic database is growing compared to the capacity of the tools and expertise that will analyze the genetic data it has. Hence, it is essential to understand that while the health industry is not the only sector experiencing big-data burden, and considering the challenges in using big data technologies are currently not in any

way unimportant, particularly given their early stages, the benefit to humanity in understanding and decoding these big genetic data makes it very important in biomedical studies and medical practices to detect disease and drugs associated with genetic variations in advanced precision medicine.

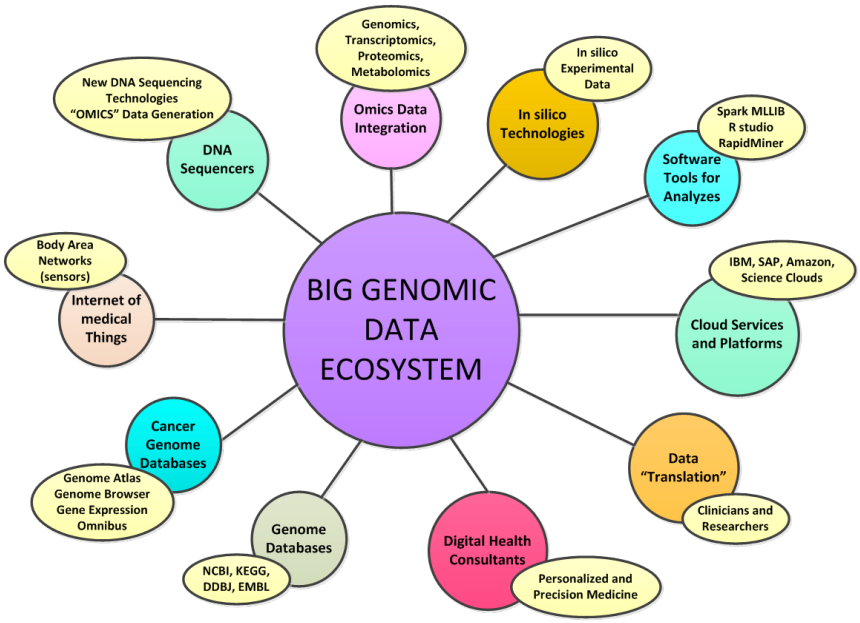

Figure 2: The Conceptual Model of the Big Genomic Data Ecosystem
(Figure adapted from an article by Plamenka Borovska, 2017)

GENOMIC DATA SCIENCE

As highlighted earlier in this survey, the study of genomics is increasingly becoming a field of study that is dominating due to the growth in the size of data and the need by the broader scientific community to utilize and store the resulting derived information effectively. While it is good to have more data from genomics, it is also crucial to highlight the widening gap that must be addressed before the broad field of genomic data science can advance in its applications. Genomics became an important field in the 1980s (Kuska et al., 1998). Since then, the field has experienced significant advancements that have made it the primary data source across all sciences. The important differences between big data in genomics and other sources can be traced to how some disciplines, such as the macromolecular structure of genomics, inherited many analytic features of genetics and natural sciences.

This survey explores the evolution of genomics and how it has become a prominent data science sub-discipline in terms of data growth and availability. Firstly, the volume, velocity, veracity, value, and variety of genomic data used as the framework for comparison to describe genomics data are examined. Secondly, this paper explores how genomics processes can be sketched for measurements, mining, modeling, and manipulation. Finally, issues relating to genomic data availability are examined.

DIFFERENCES BETWEEN GENOMIC DATA AND OTHER DATA SOURCES

Genomic data science can be compared with other disciplines in terms of volume, velocity, and variety while checking for similarities or differences in terms of other applications in data science. Another way examined genomic-data science in terms of data measurements, mining, modeling, and manipulation focusing on how physical and biological modeling of genomic data sets can be used to produce more accurate predictive models for better applications.

The Volume of Genomic Data

The volume of genomics data is one of the first things that makes genomic data different from other data sources. It serves as one of the key reasons for its prominence in data science. Reports on the volume of genomic data have not failed to report how the rate of genomic data has grown. Data growth is more significant in genomics than in other disciplines, and it is said that genomics will eventually produce more data in the coming decade [10]. This data growth has led to the problem of effective data storage and analysis. Based on the comparison, genomics data can be compared to many other data-intensive disciplines in terms of data volume. One of the ways is comparing genomic data to data generated by social sciences and comparing genomic data to other data-driven disciplines in biological sciences. Based on this comparison with social sciences, genomics data generated in genomics in terms of data volume is very large. The big-genomic ecosystem diagram below shows that there are a lot of genomic data sources, and as earlier mentioned, genomic data have grown exponentially over the past decade [9].

The Velocity of Genomic Data

A decade after the first human genome was completed in 2003 and many technologies were invented to analyze genomes due to the increased data being generated in the field. Many of these technologies were designed to increase the speed at which gene sequences were done [11]. In 2019, two significant types of sequencing technology became prominent parallel sequencers (short-read and long-read) [10]. Due to these powerful sequencers, genomic data has also entered the big-data era like other disciplines. The velocity of data can be interpreted in two ways: the speed at which data is generated and the speed at which the data is processed for use [9]. These sequencers can now sequence one human genome over 3000 times at 20-time coverage (Department of Energy Joint Genome Institute), generating data of about 50TB [10]. Hence, the capability of a bioinformatic tool to handle more extensive genomic data is needed. It is also imperative to talk about the unstructured nature of these data to indicate how important the software tools used to extract and reduce the information contained in the sequenced data need to be. New rapid sequencing technology is therefore needed in terms of the speed at which genomic data is processed.

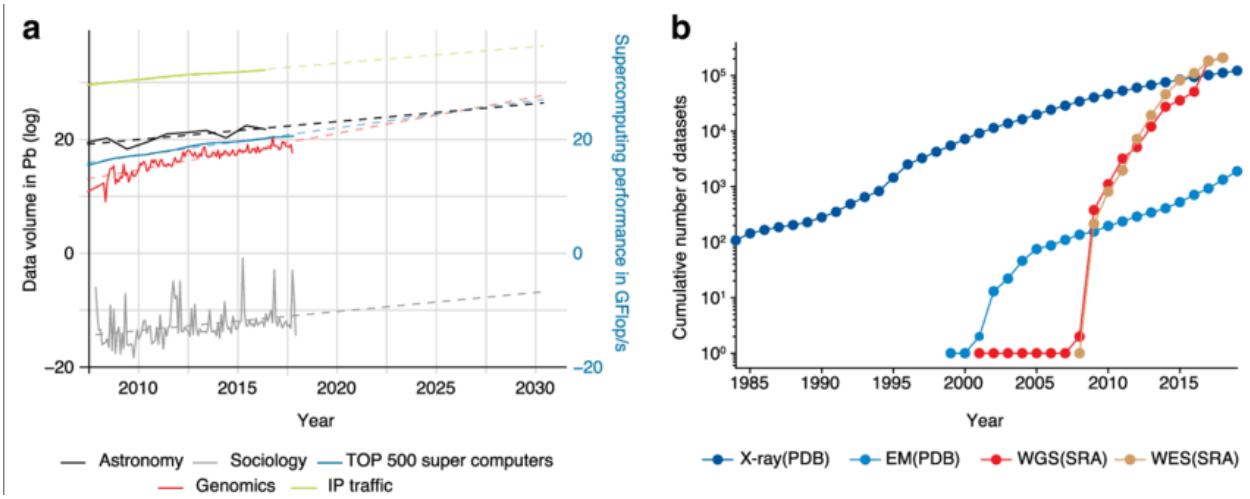

Fig 3: Data volume growth in genomics as compared to other disciplines. Fig b: Genome database growth compared to the database of other disciplines.

Variety of Data

Genomics science has not been in existence for so long, and as such, there is still no unanimity on how best to store data [1]. Research is, in fact, still ongoing on how best to store genomic data, and it has become apparent that file formats that work well with small genomes are not very practical for big-sized genomes. This issue of how genomic data is stored in many formats is another way that makes genetic data different from other data sources [1]. Genomic data can be grouped into monolithic sequencing data and phenotypic data [4]. The monolithic method of sequences has been reported to hide the variation of the set of analyses used to measure aspects of genes. However, the phenotypic data type of genomics has been reported to have many variations (examples are the simple and unstructured text from imaging data, sensors, electronic health records). It makes its data format more complicated to read and understand. Thus, more attention is being paid to standardizing and scaling these phenotypic data. The figure below illustrates the growth of the diversity of sequencing assays over time and how different sequencing methods are related to other methods. The figure below shows the number of new sequencing protocols published per year and the variety of different sequencing assays between 2006 and 2014.

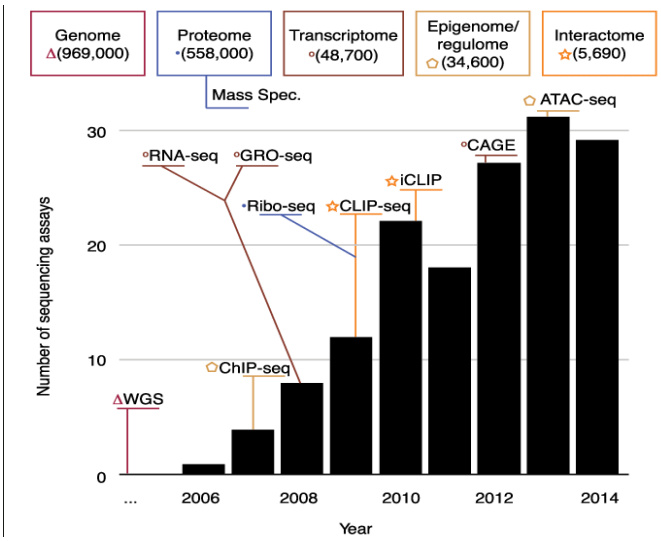

Fig 3: Variety of sequencing assays per year. (Source: research gate)

GENOMICS IN TERMS OF ITS MEASUREMENTS, MINING, MODELING, AND MANIPULATION

Science is defined as a measure to decode, describe, and make predictions on the world using distinct investigative approaches to build models. Two crucial categories differentiate science; one is the natural sciences, and the other is the social sciences. However, natural science is further divided into two main branches: biological science and physical science. In contrast to social science, data generated in natural sciences are well structured since they are generated under controlled conditions. Data generated from social sciences are not usually structured as they are generated from more subjective observations (surveys, interviews). It is essential to note that data from natural sciences mostly have underlying biological models, sometimes affecting their structure because data mining in natural sciences can be associated with mathematical modeling [6]. The overall process of genomics measuring quantity, large-scale mining, modeling, and manipulation describes a framework of what makes genomic data different [7,9]. It was reported that a reasonable way to move forward in genomics data mining and analysis could be achieved by integrating the physical & chemical mechanisms of data mining and biophysical modeling into machine learning. [6] This will help provide valuable insights that will boost data efficiency in learning new trends and patterns in genome sequencing. Data mining in genomics can estimate model parameters accurately by replacing complex parts of the model where the base theories are weak with physical models for computational efficiency. Researchers have predicted that the best way to get the best of the data in genomics will be to try to emulate how the weather forecast is done using physical-based models with large weather datasets by large-scale computing.

C O N C L U S I O N

The focus so far has been on how different genomic data is to other data-driven disciplines. The prominence of genomics amongst biological sciences was also examined. Data generated from genomics in terms of how it fits with many other areas of data science is one of the primary reasons that has made genomics data a prominent data source. Emphasis was laid on how these genomic data can benefit from data science processes and tools. Consequently, genomics' predictive power can be increased by leveraging its physical and biological aspects as used in weather forecasting.

R E F E R E N C E S

1. Joshua G. Dunn, 2014: *Categories and formats of genomics data*

2. Collins F.S., Varmus H, 2015, *A new initiative on precision medicine.*

3. Carter T.C., He M.M, 2016: *Challenges of identifying clinically actionable genetic variants for precision medicine. J. Healthc.*

4. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. 2001: *Interrelating different types of genomic data, from proteome to secretome: 'oming in on function.*

5. Eisen JA, 2012, *Badomics words and the power and peril of the ome-meme.*

6. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, et al., 2019: *Deep learning and process understanding for data-driven earth system science.*

7. Artificial intelligence alone won't solve the complexity of Earth science [Comment]. Nature. 2019.

8. Vasant Dhar, 2013: *Data Science and Prediction.*

9. Fábio C. P. Navarro, Hussein Mohsen, Chengfei Yan, Shantao Li, Mengting Gu, William Meyerson, and Mark Gerstein; 2019, *Genomics and data science: an application within an umbrella.*

10. Karen Y. He, Dongliang Ge, and Max M. He, 2017: *Big Data Analytics for Genomic Medicine*

11. Gandomi A, Haider M. 2015. *Beyond the hype: big data concepts, methods, and analytics.*