International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 558-567
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Towards Precision Medicine with Genomics using Big Data Analytics

Badugu Sobhanbabu[1]*, K.F. Bharati[2]

[1] Research Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur 515002, Andhra Pradesh, India.
* **Corresponding Author Email:** sobhanbabugec2015@gmail.com - **ORCID:** 0000-0002-1064-0584

[2] Associate Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur 515002, Andhra Pradesh, India.
Email: kfbharati.cse@jntua.ac.in - **ORCID:** 0000-0002-1809-3730

Article Info:

DOI: 10.22399/ijcesen.906
Received : 02 June 2024
Accepted : 17 January 2025

Keywords :

Big data Analytics,
Medicine,
Glycomics,
Genomics,
Machine Learning.

Abstract:

Precision medicine is considered to be the future of healthcare. It allows doctors to select treatments based on the patient's genetic information. Precision medicine is being adapted to a few typical complicated treatments like cancer at an intermediate level. As genetic information is in large volumes, Big data analytics showing a reliable promise of the modern-day health care revolution. Extremely large and continuous collection of large volumes of data like Genomics, Proteomics, Glycomics etc. is creating a challenge in analysis and interpretation, which is addressed effectively by the Big data analytics. This research work reviews and highlights the evolution of Precision medicine, Big Data Analytics and its significance in Precision medicine and related work. Also detailed the Machine learning perspectives on the Precise medicine with genomic data models along with Challenges.

1. Introduction

Precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." This method facilitates the medical practitioners and Researchers to assess the disease more precisely and lead to derive a better prevention strategies for each individual. It differs with the general practice of approving a common drug and treatment for all people in general. It takes the differences between every individual's physical responsive systems into account and provides a tailored treatment plan. Precise medicine is an emerging field of medical science targeting several diagnostic tests to conclude the medical treatments that fit best for a specific patient, though its mentioned long ago that every human is distinct. There are several considerable steps in the success of Precise medicine. A brief description of the evolution of Precise medicine is presented.

- In 1956, the "favism", the difference in individual toxicity to fava beans is identified due to metabolic deficiency of G6PD enzyme with Genetic basis
- In 1988, Renato Dulbecco revealed the mandate of Sequencing the human genome to advance in cancer Research
- In 1988, Genentech Inc. Formed a strong basis for Genomic Medicine by sequencing the entire human growth hormone. This made a revolutionized evidence about the feasibility of sequencing the human genome.
- In 1990, the famous HGP (Human Genome Project) was established, which published a first draft in 2001, followed by final version in 2003.
- In Early 1990's, the personalized treatments using the individual genome has started, but it couldn't get large focus
- In 1994, a diagnostic test was designed to predict the accuracy of rHGH replacement therapy, which is the earliest registry of a CMDx test.
- In 1998, there is an official approval of Herceptin(anti-EGFR mAb for EGFR+ breast tumors) by FDA. There onwards a huge rise in the diagnostic package, precise medicine therapies.

Since that approval and advancements in computational capabilities, Precise medicine became the most emerging field of advanced medical research. The following subtopic explains the Precise medicine evolution in the technological/computational context.

1.1 Evolution of Precision Medicine

On a Data-driven approach, Precision medicine is an emerging field, that considers the patient's pertinent genetic, medical, environmental, and behavioural information to derive tailored therapy [1-5]. The Precision medicine or personalized medicine got a revolution with the massive collection of large-scale clinical and molecular data, it elevated the expectations of biomedical research and health care [3,6]. The soul of precision medicine is to consider individual genetic profiles in every phase of health care including prevention, diagnosis, and treatment of disease[3]. The advancements in high-precision data engineering approaches that are provided with large datasets facilitate the computations required for deriving properer predictions and recommendations [7]. As the data quantity is huge, there are various challenges in analyzing and integrating such large amounts of information. To address these challenges there is a requirement for faster computational methods, more integrated processors, enhanced sensors, advanced algorithms, and methodologies cloud-based solutions which can give future directions toward precision medicine [7,8]. In the other hand, Precision medicine is not yet achieved for many clinical problems. However, the increased utilization of hypothesis-free, big data approaches assure to reach the precision medicine [9]. There are various community movements like Global Alliance for Genomics and Health (GA4GH, www.ga4gh.org), research infrastructures like ELIXIR [10], Big Data to Knowledge (BD2K) [11] and international initiatives such as the International Cancer Genome Consortium (ICGC), the International Human Epigenome Consortium (IHEC), and the International Rare Disease Consortium (IRDiRC), among others (Table 1)This review article provides an overview of Genomics, Big Data Analytics with respect to the Precise medicine with genomics, Benefits and challenges of the Big Data Analytics in the Genomics Study. The organization of the paper is as follows: Section 2 introduces the Genomics data, and its significance in Precise or Personal medicine. The Section 3 presents the Big Data and its scope with thee Genomics based Precise medicine. The Section 4 presents the Benefits of BDA in Genomics. The challenges in the Genomics-Big Data analytics is presented in Section 5. The conclusion and future works are discussed in the Section 6.

Table1. *Consortiums working on Precision medicine*

Table 1

Large international consortia focusing on Personalized Medicine

Initiative	Research focus	Link
Human Genome Diversity Project (HGDP)	General	www.hugoc.org/hgdp/
Global Network of Personal Genome Projects (PGP)	General	www.personalgenomes.org/
The Encyclopedia of DNA Elements (ENCODE)	General	www.encodeproject.org/
The NIH Roadmap Epigenomics Mapping Consortium (Roadmap)	General	www.roadmapepigenomics.org/
International Human Epigenome Consortium (IHEC)	General	http://ihec-epigenomes.org/
Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)	Cardiovascular and age-related diseases	www.chargeconsortium.com/
International Cancer Genome Consortium (ICGC)	Cancer	https://icgc.org/
The Cancer Genome Atlas (TCGA)	Cancer	https://cancergenome.nih.gov/
International Rare Disease Consortium (IRDiRC)	Rare diseases	www.irdirc.org/

2. Genomics in Precise medicine

In this section, introduction to the Genomics and its role in the Precise medicine is discussed.

2.1 Genomics is a study of Genetic material

It has processes like Sequencing, Mapping, and analysis of DNA and RNA codes of various population. The latest advancements and progress in Life Science & Healthcare is able to understand the genetic features of the people through determining entire DNA sequence. The major goal of this research is to present the relation between Genomics based Precision medicine through Big data analytics, which is also used for prevention and cure of diseases [12]. It is inevitable to adopt big data to analyze such a huge volume of data generated from sequencing, Mapping, and Analysis. Each human Genome has 20,000-25,000 genes, each consisting of 3 million base pairs. Overall, a human body's Genome data is calculated to around 100 Giga Bytes. Thus, Genomics or Genetic analysis deals with Petabytes of the data. It would lead to much more huge in volume as proceeded answer much more. The average 100 GB volume of each human body genome data makes it a challenge to conduct analysis with large sample size. The forerunner of Genomics study, "Genome Wide Association Studies (GWAS)" has conducted several studies. Currently there are more than 1600 studies that formed a unique connection between 2000 genes and more than 400 common disease symptoms. Some of the studies by the GWAS are mentioned below:

- Diabetes: Few predictive models to identify "high-risk" patients for Diabetes Type-1.
- Cancer: Classification models for guided clinical trials or highly focused cancer treatments
- Toxicity and Efficacy condition filtration with high quality information processing.

Genomics Analytics applications Genomics include several biomedical processes, each mandating huge amounts of data processing, analysis and Big Data storage and manipulations. This process can be categorized as following [12]:

1. **DNA Sequencing Library**: DNA sequencing includes separating DNA pieces based on the length through a process of electrophoresis. The sequencing system needs to maintain a huge universal library for sequencing any DNA sample (It can be even a Virun or Bacterium or a living being). As this Sequencing library has huge archives, they call for Big Data Analytics Systems.
2. **Annotation:**
   It represents addition of a description or commentary or explanation. This discription presents an explanation about each gene and its RNA product. It has the purpose of assigning a function to each gene product. Gene functions are assigned using Complex Automated Scripts' decision analysis. It is obvious that currently, some aspects has to be performed manually, but in future it can fully automated.
3. **Genomic Comparisions:** This process involves comparing billions of DNA and yielding the similarities between random sequences. This needs such a systems that are capable of Big Sequence Data and complex correlation algorithms.
4. **Genomic Visualization:** The complex correlations must be visualized along with the customizable options
5. **Synteny:** It is a process of assessing a couple of genomic regions to intuit whether its from a single ancestral genomic region. A similar Genomic comparisons based on complex statistical correlation algorithms are required for this process.

3. Big Data in Personal medicine

The Data is conceptualized in a number of different ways, including volume, velocity, diversity, value, variability, visualization, virality, and veracity, which describe the enormous amount of structured, semi-structured, and unstructured data (Figure 1) [13,14,15,16]. Health Directorate of the Directorate-General for Research and Innovation of the European Commission defined big data as "Big data in health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points [17]. Big data is widely used in the healthcare industry for many purposes like electronic health records, diagnosis reports and hospital records etc [18]. There are a decent number of measurements for sequencing of DNA, RNA, and characterizations of proteins and their properties with clinical features. For extracting useful information from a large amount of data, high-end computing solutions along with proper infrastructure are required. However, sophisticated Artificial intelligence methods including Deep learning, and cognitive computing represent their future application in healthcare in for delivering integrated solutions, predicting an outcome in big data applications [3]. Despite advances in machine learning solutions for big data, only a few have had a significant impact on clinical practice. The reasons may be a lack of validation via prospective clinical trials, inconsistent predictive performance, or difficulties interpreting complex models [18]. In working with genetic data, we should realize that the number of cases (patients) is usually very small compared to the number of genes or genetic variables measured. Because of this, the trouble is bounded through the wide variety of sufferers instead of the wide variety of variables. Consequently, the uncertainty area of the mathematical models constructed to resolve those sorts of troubles and make decisions (regressors or classifiers) could be very large, containing the set of models that are expecting the determined data with the identical error bounds. On the price feature landscape, those models are positioned in flat curvilinear valleys. The uncertainty evaluation of inverse problems and type troubles, which through definition are ill-posed, applies independently of the inverse trouble this is being solved. The noise from the information can also additionally generate spurious unphysical solutions, which makes those issues very tough to resolve. A sturdy uncertainty evaluation of the corresponding medical decision problem can then be accomplished through lowering the size of the problem [19-22]. This form of technique needs sturdy sampling strategies to recollect feasible a couple of scenarios. Data formatting and the storing of data additionally continue to be as huge challenges in the beyond years. However, the last decade has seen tremendous development in the improvement of standard genomic data codecs together with FASTQ, BAM/CRAM, and VCF files [23]. Such standardization, however, could result in incompatibility between inputs and outputs of various bioinformatics tools, or perhaps inaccurate results. As a result, imperfect standardization has enabled the sharing of genomic knowledge across institutions via federate databases such as aggregated databases such as ExAC [24] or Beacon

Network [25]. ExAC, GNOMAD, and therefore the Beacon Network databases give support in the understanding of genetic variations and distinguishing variants that are distinctive at intervals a selected ethnic group [24]. The challenges regarding downstream data formats remain, despite these successes with upstream genomic data formats. As a result, non-uniform analysis typically occurs, and re-analysis of an equivalent data exploitation completely different pipelines produces different results [26-29].

4. Big Data Applications in Genomics

Figure 1. *The Data is conceptualized in a number of different ways*

Since the completion of the sequencing of the human genome and the genetic cause of phenotyping in disease, researchers have investigated genetic markers across a large population [30-33]. This has improved efficiency by more than five orders of magnitude. Using microarrays, genome-wide analysis has been effective in evaluating population features and successfully treating complicated disorders including Crohn's disease and age-related muscle deterioration [33]. Human genome contains between 30,000 and 35,000 genes, analytics of high-throughput sequencing techniques in genomics is essentially a big data challenge [34,35]. The integration of clinical data from the genetic level to the physiological level of a human being is now being explored over a number of years [30,36]. These programmes will aid in providing each patient with individualised treatment. Fast and accurate analysis of genome-scale big data is necessary to provide suggestions in a therapeutic environment. Because investigating this big data problem requires a lot of money, time, and effort, this discipline is still in its infancy and has applications in narrowly defined emphasis areas, including cancer [37–40]. Numerous issues are covered by big data applications in genomics. Here, pathway analysis—where the functional implications of genes that are differentially expressed in experiments or gene sets of particular interest have been studied, and network reconstruction is the main focus of investigating signals obtained with high-throughput techniques to reconstruct the underlying regulatory network. The focus. These networks affect a variety of cellular functions that affect a person's physiological state [41].

4.1. Pathway Analysis

The resources for deriving the functional effects of "-omics" big data are primarily based on the statistical relationship between observed changes in gene expression and predicted functional effects. Experimental and analytical practices lead to error and batch effects [42,43]. Interpretation of functional effects should include a continuous increase in available genomic data and corresponding genetic annotations [44]. There are various tools, but there is no "gold standard" for functional pathway analysis of high-throughput genomic scale data [45-47]. The three generation methods used for path analysis [44] are described below. The first generation includes overrepresentation analysis techniques that quantify the proportion of genes involved in a certain pathway that are present among the genes that exhibit differential expression [44]. Onto-Express [45, 46], GoMiner [48], and ClueGo [46] are a few examples of first-generation tools. In the second generation, functional class scoring methods are included that take expression level variations in specific genes as well as functionally related genes into account [49]. A well-liked technique from the second generation of route analysis is GSEA [50]. The third generation of tools includes tools that are based on route topology, which are publically accessible pathway knowledge databases with precise information on the relationships of gene products, including where and how particular gene products interact with one another [44]. A third generation is demonstrated by Pathway-Analysis [51].

4.2. Reconstruction of Regulatory Networks

As an integrated operation of dynamic systems, pathway analysis does not try to understand high-throughput big data in biology [44]. Data analysis at the genome scale has been approached in a variety of ways [52-60]. Due to the broad scope of the field, the focus is on techniques for predicting networks from biological big data in this section. Systems biology uses two broad categories of network inference for big data applications: metabolic network reconstruction and gene regulatory network reconstruction [41]. A combination of different

approaches to network inference has been shown to produce better predictions [53,61]. Over the past decades, metabolic networks for reconstruction have been advanced. Through integrating genomics, transcriptomics, and proteomics high-throughput sequencing techniques, metabolic networks may be reconstructed to expand an know-how of organism-unique metabolism [52,62–68]. Constraint-primarily based totally techniques are broadly carried out to probe the genotype-phenotype dating and try to triumph over the constrained availability of kinetic constants [69, 70]. There are multitude of demanding situations in phrases of reading genome-scale facts consisting of the test and inherent organic noise, variations amongst experimental platforms, and connecting gene expression to response flux utilized in constraint-based techniques [71,72].

Available reconstructed metabolic networks encompass Recon 1 [62], Recon 2 [52], SEED [64], IOMA [66], and MADE [73]. Recon 2 (a development over Recon 1) is a model to symbolize human metabolism and contains 7,440 reactions related to 5,063 metabolites. Recon 2 has been accelerated to account for recognized drugs for drug target prediction studies [74-81] and to examine off-target consequences of drugs [74]. Reconstruction of gene regulatory networks from gene expression statistics is another emerging field. Network inference strategies may be split into five classes primarily based on the underlying version in every case: Regression, Mutual records, Correlation, Boolean regulatory networks, and other strategies [53].

More than 30 inference strategies have been assessed after DREAM5 task in 2010 [53]. Performance various inside every class and there has been no class located to be continuously higher than the others. Different strategies make use of distinctive records to be utilized in experiments which may be withinside the shape of time series, drug perturbation experiments, gene knockouts, and mixtures of experimental conditions.

A tree based method (the use of ensembles of regression trees) [75] and two-manner ANOVA (evaluation of variance) method [76] gave the best overall performance in a current DREAM task [61]. Nodes and sets of nodes are governed by boolean regulatory networks [41], which are special cases of discrete dynamical models. Using Boolean operations on the states of other nodes in the network, the state of each node or set of nodes can be determined by determining the actual state of each node or set of nodes [54].

By using prior information, boolean networks may be able to reduce the number of false positives (i.e., when a condition appears to be satisfied while in reality it is not) although they can be extremely useful when the amount of quantitative data is small [41, 54]. When there are many nodes in a network, Boolean networks are prohibitively expensive. There are more global states than entities, which is due to the exponential increase in the number of entities [41]. Utilizing clustering to reduce the size of the problem is one way to get around this bottleneck. For instance, Martin et al. [79] used clustering methods to divide a microarray gene expression dataset containing 34,000 probes in 23 sets of metagenes. For two separate immunology microarray datasets, our Boolean model effectively represented network dynamics. ODEs may be used to model the dynamics of a gene regulatory network [56–59]. This method has been used to identify the yeast regulatory network [56].

The regulatory network which molecular biologists have used experiments to define was successfully captured by the study. It takes a lot of computing power to reconstruct a gene regulation network on a Genome scale system as a dynamic model [41]. To deal with this issue, a parallelizable dynamical ODE model has been created[80]. It drastically cuts down on computation time[80]. Exploring nearly a billion potential connections is necessary to identify connections in the regulatory network for a challenge the size of the human genome, which contains 30,000–35,000 genes [34, 35]. The cardiogenic gene regulation network of the mammalian heart has been rebuilt using the dynamical ODE model [59]. Table 2 lists many techniques and toolkits along with the applications they can be used for.

Table 2. *Famous projects methods details with their applications*

Projects	**Analysis category**	**Applications**
Onto-Express [45,46]	Pathway analysis.	Breast cancer[46]
GoMiner [48]	Pathway analysis.	Pancreatic cancer [44]
ClueGo [49]	Pathway analysis.	Colorectal tumors [46]
GSEA [50]	Pathway analysis.	Diabetes [48]
Pathway-Express [51]	Pathway analysis.	Leukemia [50]
Recon 2 [52]	Reconstruction of metabolic networks	Drug target prediction studies [52]
Boolean methods [41,53,54]	Reconstruction of gene regulatory networks	Cardiac differentiation [55]
ODE models [56-59]	Reconstruction of gene regulatory networks	Cardiac development [59]

4.3 Genome Sequencing Cost reduction

The cost of sequencing the human genome was around $100 million a few years ago [12]. It may be less than $50 million today. The downward trend will only continue in the coming years as Big Data adoption grows. In recent years, healthcare researchers have also worked to reduce the cost of genome sequencing and make it more accessible for everyone. Today, an individual human genome sequencing costs only around $5000[12].

4.4 Time Saving

With a traditional setup and a lot of data stored in databases, extract, transform, load (ETL) tests would take a long time. There is no ETL with Big Data solutions like Hadoop. Thus, data analysis is relatively quick, which would save a lot of time. Spark and Python are being widely adopted tools enabling a steady handshake and an easier time to market solution to work on this Big Data with far greater results."

4.5 Better Analysis

Hadoop like Big Data systems permits us to conduct analysis that is not possible in a regular machine intelligence tool, it will not even work for an traditional SQL relational type setup.

5. Challenges in Precise medicine development

This section lists some of the major challenges that Precision Medicine faces, especially with Machine Intelligence and Big Data. Data Storage Costs Big data creation and accessing generate major constsaint for storage [12], transfer and security of information.

Now, its is less expensive to generate when comapred to stroring that data. For example, the NCBI, a forerunner of Big Data application in biomedical science since 1988, couldn't be able to arrive at a comprehensive, safe and less cost solution to the data storing constraint.

Adoption of Big Data Techniques

There exists a difference among the implementation and envisioning of Big data in Data sciences.

To achieve a proper solution, the Big data problems need to be converted to reduced dimensions and achievable data problems [12]. The major constraint inn adopting the Big data is the cost-to-benefit analysis, that highlights and converts a workable solution to a business problem and commercializes with a quantifiable Rol churn.

Large Initial Investment

It require huge captial investment for Big data management, which may be not possible for small organizations or laboratories [12]. It is one of the limitation for conducting vide biomedical research with big data.

Big Data Transfer

One of the biggest problems is data transfer design. It is expensive and complex to design the big data transfer protocols and real-time execution. Currently, the data transfer is managed with physical external hard disks, which is indeed not a good option for future analysis. An alternative solution is to use Biotorrents of data transfer, which were initially developed for facilitating large data transfer through the internet for biomedical research.

Security & Privacy

Retaining the confidentiality of the data is an ethical concern [12]. Solving this problem is an expensive matter. Advanced cryptographic algorithms for encryption of the data is an essential step to preserve security and privacy. Sophisticated distributed data systems like Blockchain [86] are required and may be used in the future [82-87].

6. Conclusion and Future Scope:

The future of Precision medicine lies in the hands of Big data analytics that can hold a grip on managing structured as well as unstructured data sources and will play a vital role in how the healthcare sector will be practiced [82]. For prediction and decision making some analytics are already in practice by some healthcare professionals and organizations. The major focus I in three areas as Physiological signal processing, Image analysis and Genomic data processing. As medical data including genomic data is growing exponentially, which mandating the computational scientists to proceed with innovative solutions to analyze huge volumes of data in as much time as possible. Such trend adoption of big data practice is being observed from the healthcare professionals, this adoption is growing steadily with h development of very imaginative and incredible systems that produce precision medicine that is saving many peoples life. It is not exaggerating that Precision medicine gives a second life for the critical staged patients. By combining physiological data with high-throughput "-omics" techniques, we can create a detailed model of the human body that can improve both contribute to the creation of blood-based diagnostic tools and further our understanding of illness states [83–85]. Medical image analysis, signal processing, and the integration of physiological and "-omics" data provide comparable

potential and obstacles when using diverse organised and unstructured big data sources. Big data Analytics is used for different application [88,89].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Hassan M, Awan FM, Naz A, deAndrés-Galiana EJ, Alvarez O, Cernea A, Fernández-Brillet L, Fernández-Martínez JL, Kloczkowski A. (2022). Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. *Int J Mol Sci.* 23(9):4645. doi: 10.3390/ijms23094645.

[2] Hulsen Tim, Jamuar Saumya S., Moody Alan R., Karnes Jason H., Varga Orsolya, Hedensted Stine, Spreafico Roberto, Hafler David A., McKinney Eoin F. (2019). From Big Data to Precision Medicine, *Frontiers in Medicine*, DOI=10.3389/fmed.2019.00034

[3] Cirillo, D.; Valencia, A. (2019). Big data analytics for personalized medicine. *Curr. Opin. Biotechnol*. 58, 161–167.

[4]Ginsburg, G.S.; Willard, H.F. (2019). Genomic and personalized medicine: Foundations and applications. *Transl. Res*. 2009, 154, 277–287.

[5]Naqvi, M.R.; Jaffar, M.A.; Aslam, M.; Shahzad, S.K.; Iqbal, M.W.; Farooq, A. (2020). Importance of big data in precision and personalized medicine. In *Proceedings of the 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 30 July 2020; pp. 1–6

[6]Iriart, J.A.B. (2019). Precision medicine/personalized medicine: A critical analysis of movements in the transformation of biomedicine in the early 21st century. *Cadernos. Cad. De Saúde Publica* 35.

[7]Beckmann, J.S.; Lew, D. (2016) Reconciling evidence-based medicine and precision medicine in the era of big data: Challenges and opportunities. *Genome Med*. 8, 1–11.

[8]Espinal-Enríquez, J.; Mejía-Pedroza, R.; Hernández-Lemus, E. (2017) Computational approaches in precision medicine. *Progress and Challenges in Precision Medicine;* pp. 233–250

[9]Hulsen, T.; Jamuar, S.; Moody, A.; Karnes, J.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.; McKinney, E. (2019). From big data to precision medicine. *Front. Med.* 6, 34.

[10]Durinx, Christine, Jo McEntyre, Ron Appel, Rolf Apweiler, Mary Barlow, Niklas Blomberg, Chuck Cook et al. (2016) Identifying ELIXIR core data resources. *F1000Research* 5.

[11]Ronald Margolis, Leslie Derr, Michelle Dunn, Michael Huerta, Jennie Larkin, Jerry Sheehan, Mark Guyer, Eric D Green, (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data, *Journal of the American Medical Informatics Association*, 21(6);957–958, https://doi.org/10.1136/amiajnl-2014-002974

[12]Genomics Reimagined by Big Data & Analytics – 3AI, October 22, 2020, Genomics Reimagined by Big Data & Analytics – 3AI

[13]Bibault, J.-E. (2020). Real-life clinical data mining: Generating hypotheses for evidence-based medicine. *Ann. Transl. Med.* 8, 69.

[14]Normandeau, K. (2013). Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. *Inside Big Data* https: //insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ (accessed on 18 January 2022).

[15]Gandomi, A.; Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* 35, 137–144.

[16]Diebold, F.X.; Cheng, X.; Diebold, S.; Foster, D.; Halperin, M.; Lohr, S.; Mashey, J.; Nickolas, T.; Pai, M.; Pospiech, M. (2012). A Personal Perspective on the Origin (s) and Development of "Big Data": *The Phenomenon, the Term, and the Discipline*.

[17]Auffray, C.; Balling, R.; Barroso, I.; Bencze, L.; Benson, M.; Bergeron, J.; Bernal-Delgado, E.; Blomberg, N.; Bock, C.; Conesa, A. (2016). Making sense of big data in health research: Towards an EU action plan. *Genome Med.* 8, 1–13.

[18]Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *J. Big Data* 6, 54.

[19]Fernandez Martinez, J.L.; Fernandez Muniz, M.Z.; Tompkins, M.J. (2012). On the topography of the cost functional in linear and nonlinear inverse problems. *Geophysics* 77, W1–W15.

[20]Fernández-Martínez, J.L.; Fernández-Muñiz, Z.; Pallero, J.; Pedruelo-González, L.M. (2013). From Bayes to Tarantola: New insights to understand uncertainty in inverse problems. *J. Appl. Geophys.* 98, 62–72.

[21]Fernández-Martínez, J.L.; Pallero, J.; Fernández-Muñiz, Z.; Pedruelo-González, L.M. (2014) The effect of noise and Tikhonov's regularization in inverse

problems. Part I: The linear case. *J. Appl. Geophys.* 108, 176–185.

[22]Fernández-Martínez, J.L.; Pallero, J.; Fernández-Muñiz, Z.; Pedruelo-González, L.M. (2014). The effect of noise and Tikhonov's regularization in inverse problems. Part II: The nonlinear case. *J. Appl. Geophys.* 108, 186–193.

[23]Zhang, H. (2016) Overview of sequence data formats. In *Statistical Genomics; Springer: Berlin/Heidelberg, Germany,* pp. 3–17.

[24]Lek, M.; Karczewski, K.; Minikel, E.; Samocha, K.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.; Ware, J.; Hill, A.; Cummings, B.; et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536, 285–291.

[25] Peter Goodhand et al. The Global Alliance for Genomics and Health. A federated ecosystem for sharing genomic, clinical data. *Science*, 352, 1278–1280. DOI: 10.1126/science.aaf6162

[26]Wenger, A.M.; Guturu, H.; Bernstein, J.A.; Bejerano, G. (2017). Systematic reanalysis of clinical exome data yields additional diagnoses: Implications for providers. *Genet. Med*., 19,209–214.

[27]Wright, C.F.; McRae, J.F.; Clayton, S.; Gallone, G.; Aitken, S.; FitzGerald, T.W.; Jones, P.; Prigmore, E.; Rajan, D.; Lord, J. (2018). Making new genetic diagnoses with old data: Iterative reanalysis and reporting from genome-wide data in 1133 families with developmental disorders. *Genet. Med.* 20, 1216–1223.

[28]Anuradha, T., Lakshmi Surekha, T., Nuthakki, P., Domathoti, B., Ghorai, G., Shami, F.A. (2022) Graph theory algorithms of Hamiltonian cycle from quasi-spanning tree and domination based on vizing conjecture. *J. Math.* https://doi.org/10.1155/2022/1618498

[29]Sekhar, J. N. C., Domathoti, B., & Santibanez Gonzalez, E. D. R. (2023). Prediction of Battery Remaining Useful Life Using Machine Learning Algorithms. *Sustainability*, 15(21), 15283. https://doi.org/10.3390/su152115283

[30]L. Hood and N. D. Price, (2014). Demystifying disease, democratizing health care, *Science Translational Medicine*, 6,225, Article ID 225ed5, 2014.

[31]J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing, *Nature Reviews Genetics*, 12(7);499–510.

[32]T. J. Treangen and S. L. Salzberg, (2012). Repetitive DNA and next generation sequencing: computational challenges and solutions, *Nature Reviews Genetics*, 13(1);36–46.

[33]D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, (2013). The next-generation sequencing revolution and its impact on genomics, *Cell*, 155(1);27–38.

[34]E. S. Lander, L. M. Linton, B. Birren et al., (2001). Initial sequencing and analysis of the human genome, *Nature*, 409(6822);860–921.

[35]R. Drmanac, A. B. Sparks, M. J. Callow et al., (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, *Science*, 327(5961);78–81.

[36]R. Chen, G. I. Mias, J. Li-Pook-Than et al., (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes, *Cell*, 148(6);1293–1307,

[37]Institute of Medicine, Informatics Needs and Challenges in Cancer Research: Workshop Summary, *The National Academies Press*, Washington, DC, USA, 2012.

[38]E. M. van Allen, N. Wagle, and M. A. Levy, (2013) Clinical analysis and interpretation of cancer genome data, *Journal of Clinical Oncology*, 31(15);1825–1833.

[39]A. Tabchy, C. X. Ma, R. Bose, and M. J. Ellis, (2013) Incorporating genomics into breast cancer clinical trials and care, *Clinical Cancer Research*, 19(23);6371–6379.

[40]F. Andre, E. Mardis, M. Salm, J. C. Soria, L. L. Siu, and C. Swanton, (2014). Prioritizing targets for precision cancer medicine, *Annals of Oncology*, 25(12);2295–2303.

[41]G. Karlebach and R. Shamir, (2008). Modelling and analysis of gene regulatory networks, *Nature Reviews Molecular Cell Biology*, 9(10);770–780.

[42]J. Loven, D. A. Orlando, A. A. Sigova et al., (2012). Revisiting global ´ gene expression analysis, *Cell*, 151(3);476–482.

[43]J. T. Leek, R. B. Scharpf, H. C. Bravo et al., (2010) Tackling the widespread and critical impact of batch effects in high throughput data, *Nature Reviews Genetics*, 11(10);733–739.

[44]P. Khatri, M. Sirota, and A. J. Butte, (2012). Ten years of pathway analysis: current approaches and outstanding challenges, *PLoS Computational Biology*, 8(2);e1002375.

[45]P. Khatri, S. Draghici, G. C. Ostermeier, and S. A. Krawetz, (2001). Profiling gene expression using Onto-Express, *Genomics*, 79(2);266–270.

[46]S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, (2003). Global functional profiling of gene expression, *Genomics*, 81(2)98–104.

[47]D. W. Huang, B. T. Sherman, and R. A. Lempicki, (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Research*, 37(1);1–13.

[48]B. R. Zeeberg, W. Feng, G. Wang et al., (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data, *Genome Biology*, 4(4);R28,

[49]G. Bindea, B. Mlecnik, H. Hackl et al. (2009)., Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics*, 25(8);1091–1093

[50]A. Subramanian, P. Tamayo, V. K. Mootha et al., (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, 102(43);15545–15550.

[51]S. Draghici, P. Khatri, A. L. Tarca et al., (2007). A systems biology approach for pathway level analysis, *Genome Research*, 17(10);1537–1545.

[52]I. Thiele, N. Swainston, R. M. T. Fleming et al., (2013). A communitydriven global reconstruction of human metabolism, *Nature Biotechnology*, 31(5);419–425.

[53]D. Marbach, J. C. Costello, R. Kuffner et al.,(2012). Wisdom of crowds ¨ for robust gene network inference, *Nature Methods*, 9(8);796–804.

[54]R.-S.Wang, A. Saadatpour, and R. Albert, (2012). Boolean modeling in systems biology: an overview of methodology and applications, *Physical Biology*, 9(5);055001.

[55]W. Gong, N. Koyano-Nakagawa, T. Li, and D. J. Garry, (2015). Inferring dynamic gene regulatory networks in cardiac differentiation through the integration of multi-dimensional data, *BMC Bioinformatics*, 16(1);74.

[56]K. C. Chen, L. Calzone, A. Csikasz-Nagy, F. R. Cross, B. Novak, and J. J. Tyson, (2004). Integrative analysis of cell cycle control in budding yeast, *Molecular Biology of the Cell*, 15(8);3841–3862.

[57]S. Kimura, K. Ide, A. Kashihara et al., (2005). Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, *Bioinformatics*, 21(7);1154–1163.

[58]J. Gebert, N. Radde, and G.-W. Weber, (2007). Modeling gene regulatory networks with piecewise linear differential equations, *European Journal of Operational Research*, 181(3);1148–1165.

[59]J. N. Bazil, K. D. Stamm, X. Li et al., (2014). The inferred cardiogenic gene regulatory network in the mammalian heart, *PLoS ONE*, 9(6);e100842.

[60] B. O. Palsson, Systems Biology, Cambridge University Press, 2006.

[61]D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, (2010). Revealing strengths and weaknesses of methods for gene network inference, *Proceedings of the National Academy of Sciences of the United States of America*, 107(14)6286–6291.

[62]N. C. Duarte, S. A. Becker, N. Jamshidi et al.,(2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data, *Proceedings of the National Academy of Sciences of the United States of America*, 104(6);1777–1782.

[63]K. Raman and N. Chandra, (2009). Flux balance analysis of biological systems: applications and challenges, *Briefings in Bioinformatics*, 10(4);435–449.

[64]C. S. Henry, M. Dejongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models, *Nature Biotechnology*, 28(9);977–982.

[65]K. Radrich, Y. Tsuruoka, P. Dobson et al., (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks, *BMC Systems Biology*, 4,114.

[66]K. Yizhak, T. Benyamini, W. Liebermeister, E. Ruppin, and T. Shlomi, (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model, *Bioinformatics*, 26(12);i255–i260.

[67]C. R. Haggart, J. A. Bartell, J. J. Saucerman, and J. A. Papin, (2011). Whole-genome metabolic network reconstruction and constraint-based modeling, in *Methods in Systems Biology, M. Verma, D. Jameson, and H. V. Westerhoff, Eds., vol. 500 of Methods in Enzymology*, chapter 21, pp. 411–433, Academic Press.

[68]D. McCloskey, B. Ø. Palsson, and A. M. Feist, (2013). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli, *Molecular Systems Biology*, 9,661.

[69]E. P. Gianchandani, A. K. Chavali, and J. A. Papin, (2010). The application of flux balance analysis in systems biology, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(3);372–382.

[70]N. E. Lewis, H. Nagarajan, and B. O. Palsson, (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods, *Nature Reviews Microbiology*, 10(4);291–305.

[71]W. Zhang, F. Li, and L. Nie, (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies, *Microbiology*, 156(2);287–301.

[72]A. S. Blazier and J. A. Papin, (2012). Integration of expression data in genome-scale metabolic network reconstructions, *Frontiers in Physiology*, 3,299.

[73]P. A. Jensen and J. A. Papin, (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding, *Bioinformatics*, 27(4);541–547.

[74]R. L. Chang, L. Xie, L. Xie, P. E. Bourne, and B. Ø. Palsson, (2010). Drug off-target effects predicted using structural analysis in the context of a metabolic network model, *PLoS Computational Biology*, 6(9);e1000938.

[75]V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, (2010). Inferring regulatory networks from expression data using tree-based methods, *PLoS ONE*, 5(9);e12776.

[76]R. Kuffner, T. Petri, P. Tavakkolkhah, L. Windhager, and R. (2012). Zimmer, "Inferring gene regulatory networks by ANOVA, *Bioinformatics*, 28(10);1376–1382.

[77]R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and G. Stolovitzky, (2011). Crowdsourcing network inference: the dream predictive signaling network challenge, *Science Signaling*, 4;189.

[78]T. Saithong, S. Bumee, C. Liamwirat, and A. Meechai, (2012). Analysis and practical guideline of constraint-based boolean method in genetic network inference, *PLoS ONE*, 7(1);e30232.

[79]S. Martin, Z. Zhang, A. Martino, and J.-L. Faulon, (2007). Boolean dynamics of genetic regulatory networks inferred from microarray time series data, *Bioinformatics*, 23(7);866–874.

[80]J. N. Bazil, F. Qi, and D. A. Beard, (2011). A parallel algorithm for reverse engineering of biological networks, *Integrative Biology*, 3(12);1215–1223.

[81]O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, (2011). Predicting selective drug targets in cancer through metabolic networks, *Molecular Systems Biology*, 7(1).

[82]Atmika Sharma, (2016). Big Data: A Healthcare Revolution, *International Research Journal of Engineering and Technology (IRJET)*, 03(08).

[83]L. Hood and S. H. Friend, (2011). Predictive, personalized, preventive, participatory (P4) cancer medicine, *Nature Reviews Clinical Oncology*, 8(3);184–187.

[84]L. Hood and M. Flores, (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory, *New Biotechnology*, 29(6);613–624.

[85]L. Hood and N. D. Price, (2014). Demystifying disease, democratizing health care, *Science Translational Medicine*, 6(225);225ed5.

[86]Anil, G. R. and Salman Abdul Moiz. (2019). Blockchain Enabled Smart Learning Environment Framework. *Learning and Analytics in Intelligent Systems*.

[87]Van Loenen, B.; Kulk, S.; Ploeger, H. (2016). Data protection legislation: A very hungry caterpillar: The case of mapping data in the European Union. *Gov. Inf. Q*. 33, 338–345.

[88]M. Venkateswarlu, K. Thilagam, R. Pushpavalli, B. Buvaneswari, Sachin Harne, & Tatiraju.V.Rajani Kanth. (2024). Exploring Deep Computational Intelligence Approaches for Enhanced Predictive Modeling in Big Data Environments. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.676

[89]K.S. Praveenkumar, & R. Gunasundari. (2025). Optimizing Type II Diabetes Prediction Through Hybrid Big Data Analytics and H-SMOTE Tree Methodology. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.727