Data for Genomics in Healthcare: The Potential of Next-Generation Sequencing in Developing Precision Medicine

Dafa Dwi Antara¹, Michael Yahezkiel Lie¹, Muhammad Darisza¹, Nicola Kertapati¹, Ivan Valentino¹, Bharuno Mahesworo², Ika Nurlaila²,³, Haryono Soeparno¹,²

¹Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
²Bioinformatics & Data Science Research Center, Bina Nusantara University Jakarta, Indonesia 11480
³Information Systems Department, BINUS Online Learning, Bina Nusantara University Jakarta, Indonesia 11480

bharuno.mahesworo@binus.edu

Abstract. Recent advances in genomics on healthcare technology have opened the path towards more precise treatments by analysing patients’ genome using a high throughput DNA sequencing technology called Next Generation Sequencing (NGS). The information obtained is further used to determine a more effective treatment path for patients based on their genetic profiles. It can also be employed to predict the risk of patients developing certain diseases. Despite particular difficulties it might create in the first phase of settling the integrative frameworks down, a successful combination of NGS and Big Data technology would offer a huge improvement in terms of the effectiveness of treatments and medicine administration for the patients. This paper elucidates a general overview of how NGS-based DNA sequencing makes precision medicine possible and describes the use of Big Data for genomics in healthcare-wise.

1. Introduction

Genomics studies have recently faced a rapid development following the much lower cost spent on genomics information retrieved from next-generation sequencing (NGS) [1]. NGS is a high throughput technology that is employed with an aim at sequencing millions of small fragments of human’s DNA [2] to develop personalized treatment termed precision medicines [3]. Precision medicine is defined as pre-tailored medical treatment based on their unique genetic characteristics hence their susceptibility upon specific treatment might differ across subpopulation and could be defined by subpopulation [4]. The negative impact of the production of pharmaceutical products on the natural environment is well known. Thus, using tailored medicine or treatment reduce the negative impact on the environment caused by the pharmaceutical industry. It allows doctors to predict the risks of patients to develop certain diseases by employing NGS to perform a complete genome resequencing [5] rendering it feasible to determine existing mutations in their genomes or compare the patient’s genome to other patients with similar sequences [6]. There are various existing NGS platforms, including Roche GS-FLX 454 Genome Sequencer, the ABI SOLiD platform, the Polonator G.007, the Helicos HeliScope, and the Illumina/Solexa Genome Analyzer [7].
Previous studies on the implementation of clinical data held in electronic medical records (EMRs) which is well-known as electronic health records (EHRs) suggested that these could be tailored in genomic studies [8] through which patients would be benefitted by getting regiments based on their particular genome structures thus was expected to be more effective as they target specific affected sequences for genomic research purposes [9]. Despite the resourcefulness of genomic data, its implications in medical practices demand relatively sophisticated infrastructure to accommodate its gigantic data size and complexity [10]. The DNA of a human comprises of 3 billion base pairs with a genome having approximately 100 gigabytes of data [11]. To be able to accommodate such a big data, adequate infrastructure is of absolute requirement. The phrase of Big Data refers to collections of massive datasets which causes them unfit to be stored and processed by conventional data storing and processing methods [12]. In recent years, Big Data has been applied to various fields such as medical image processing, neurology, proteomics, genomics, and many others [13], [14].

Therefore, Big Data infrastructure is vital in orchestrating precision medicine. The present study aims at disseminating general understandings on tailoring NGS and Big Data in the healthcare-wise for the advancement of precision medicine as well as elaborating some aspects that might hinder the progressions. The result of this study paves future research or project that develop or implement big data infrastructure toward NGS data.

2. Literature Review

2.1. Next-Generation Sequencing
The term of next-generation sequencing defines the high-throughput sequencing methods that allow millions or even trillions of observations being done in parallel during a single instrument run [2], [15]. Over the last thirty years, DNA sequencing technologies and applications have undergone tremendous development, becoming the force that moves the field of genomics with a vast amount of genome data and various research areas and multiple applications [15]–[17]. Examples of NGS technologies include the Roche 454 system, AB SOLiD system and Illumina Genome Analyzer system [7]. The applications of NGS technologies hitherto are mainly in the healthcare-wise where the determination of genetic-susceptibility upon particular diseases is facilitated [5], [6], [18].

2.2. Precision Medicine
The term precision medicine refers to treatments that are pre-designed specifically according to pre-defined characteristics of an individual which include genetic-, phenotype, or psycho-social-profiles. Not only can the precision medicine improve the effectiveness and efficiency of the treatments. It is also believed to lower chances of experiencing deleterious or deteriorating consequences for the overall health condition [19], [20]. The technology improvement in NGS and the implementation of EHRs will significantly improve the development of precision medicine in the actual clinical practice. As aforementioned, genomics information generated by NGS technologies play a vital role in supporting precision medicine which is obviously a patient-centralized platform[3], [9], [18], [21].

2.3. Big Data Analytics in Healthcare
Big Data technology has been used in many fields, one of which being in the healthcare-wise [6], [10]. The massive volume of heterogeneous medical data causes it tricky to manage by using conventional means [10]. With the use of Big Data technology however, this issue can be resolved. Storing, handling, and processing medical data or EHRs can be done using Big Data technology [8], [10], [21]. In this regard, Big Data technology is critical in the alluded field. Diverse Big Data frameworks with a focus on processing medical data [22]–[24] have been proposed over the last 5 years, indicating a strong connection between the two. By integrating NGS and Big Data technologies, precision medicine is no longer an unsolved issue to formulate.

2.4. Previous Studies
Gullapalli et al summarized the requirements and hurdles for a successful implementation of sequencing systems that use Next-Generation Sequencing methods [25]. Their study suggested that the development of NGS methods has enabled the potential characterization of the patient’s genome as the foundation of appropriate instrument platforms development, bioinformatics tools, and the evolution of the workforces [25]. When succeeds, these systems will provide valuable diagnostics information which in turn will responsiveness of the pertinent patients toward the recommended regimens as these have been pre-formulated based on their unique genetic characteristics in combination with their EHR information [8], [10], [21].

Jay Shendure and Hanlee Ji reviewed next-generation DNA sequencing platforms by explaining their work process, strengths and limits [26]. They also discussed emerging applications and other developments in the field of DNA sequencing technology. Different next-generation DNA sequencing technologies also differ in served strength and limitations which in turn influence the cost and reading time. NGS can be applied in myriads of things. One example is a complete genome resequencing to discover mutations in individual human genomes.

Thorvaldsdóttir H. et al described the Integrative Genomics Tool (IGV) in their study, a high-performance data visualization tool with a focus on genomic studies [22]. The tool can cope with large and heterogeneous datasets to support next-generation sequencing data and integration of clinical data. The IGV is a tool with a few useful features and utilities, such as the ability to load local and remote datasets that can be used by researchers to assist in genomic data visualization. It is available for free.

J. Larry Jameson and Dan L. Longo summarized the forces accelerating precision medicine, the challenges of implementing precision medicine, and the implications of precision medicine for clinical practice in their study [27]. They found that genetics and next-generation DNA sequencing methods have the greatest effect in expanding the scope of precision medicine. Although implementing precision medicine would enable patients to get improved clinical outcomes and minimize side effects from ineffective treatments, it is without its challenges, such as the complexity of the data involved and concerns from payers regarding the actual effectiveness of this innovation.

Blagoj Ristevski and Ming Chen surveyed the use of Big Data analytics in medicine and healthcare [28]. They found the challenges faced regarding the integration of Big Data in healthcare. The ability to explore, process, and analyze large amounts of complex heterogeneous data such as biomedical data, social media data, and EHRs makes Big Data analytics a vital component in healthcare. By using frameworks that can handle Big Data, it enables the development of medical applications that applies data mining techniques to discover hidden patterns and new knowledge.

O’Driscoll A. et al discussed cloud computing and big data technologies as a potential solution in storing, handling, and processing the ever-growing biological data, particularly regarding sequence data [29]. They also discussed that implementing Big Data technologies, not only storing data but also processing data promptly is challenging. There have been numerous leading projects that utilize the Apache Hadoop platform to analyze and process large biological datasets, but there is still much to discover. Big Data certainly has the potential to revolutionize healthcare in many ways.

Kaur P. et al provide a brief introduction to Big Data and its applications in their study [23]. The study based on existing researches done in 2010 up until 2016 with the main intention of proposing an optimized and secure big data healthcare framework. The proposed framework model contains four layers: data source layer, data storage layer, data security and privacy layer, and machine learning-based application layer. This model emphasizes the optimization and security of patients’ data.

Zhang Y. et al proposed a cyber-physical system for patient-centric healthcare applications and services called Health-CPS that is built on cloud and Big Data analytics technologies [24]. This research also aims to show that those technologies (cloud and Big Data) can be used to enhance the performance of the healthcare system. The proposed system is assisted by cloud and Big Data, with the architecture consisting of three layers: data collection layer, data management layer, and application service layer. The proposed system can be used for at least four different groups of applications: statistic-based, monitoring-based, knowledge-based, and prediction-based.
3. Materials and Methods

3.1. Research Questions
The purpose of this research is to provide insights for the readers on how NGS and Big Data can be tailored to support precision medicine. To arrange the results systematically, 3 questions are taken as a guide:

- What role does NGS have in developing precision medicine?
- What are the challenges faced in the development of precision medicine?
- How can Big Data technologies be used to support precision medicine?

3.2. Data Collection Method
Our data collection is by academic databases-derived literature reviews. A total of 43 original research papers were gathered using some relevant keywords, being that: Next Generation Sequencing and Big Data in Healthcare. The selected papers were pre-screened to evaluate the respective relevancy toward the focus of the present study, thus we processed only eight papers, eventually, for they demonstrated tuning elaborations on NGS and implementation of Big Data in healthcare.

4. Results and Discussion
Genetics and next-generation DNA sequencing methods are tremendously impactful for expanding the scope of precision medicine to achieve clinical outlook improvements. The development of diverse NGS methods has allowed the characterization and complete genome resequencing of a patient’s genome to discover mutations. Therefore, early detection and prediction can be based on a strong basis where uniqueness and population/subpopulation identity are considered. Consequently, regiments can be pre-tailored more effectively as the affected gene and health-related conditions have been previously captured and defined.

Despite all those mentioned desired outcomes of integrated NGS data and medical records, precision medicine is deemed to be challenging not only in term of minimum infrastructure demanded but also in term of availability of massive targeted gene sequences which represent diverse population either those of typically represented or under-represented populations. This seems to be hardly fulfilled whenever the priority and commitment are not set and are not paid with adequate attention from relevant stakeholders who, at the end of this development, will be the executor of the strategy in the actual field.

According to a study conducted by Jameson, et al. [27], people’s scepticism is another challenge which is determinative toward the effectiveness of the applied precision medicine. To achieve success in precision medicine, a fit platform that can be used to store the massive size of medical data and process those ever-growing medical data is required.

Big Data analytics has a power of exploration, processing and analysis of large size heterogeneous data such as biomedical data, social media data, and EHRs. Big data has been applied in several applications namely statistic-based, monitoring-based, knowledge-based, and prediction-based. The use of certain frameworks that can cope with Big Data, for instance Apache Hadoop, facilitates the development of medical applications that applies data mining techniques to discover hidden patterns and reveal new knowledge. There have been a number of proposed Big Data frameworks to handle medical data reported in some previous studies, such as a model proposed by Kaur, P. et al. [23] that combines Big Data technology with machine learning and a model proposed by Zhang, Y. et al. [24] that uses Big Data technologies combined with cloud computing. There are also visualization tools that enable the viewing of genomics data, called the IGV [17], which can also be employed to visualize medical data such as EHRs obtained from the data mining. The successful integration of Big Data technologies is pivotal for precision medicine implementation.

5. Conclusion
As the development of genomic-based appliances escalates rapidly, abundant discoveries about genomics information emerge. This allows us to formulate precision medicine as a new paradigm in the
healthcare-wise although a few basic requirements which are resources-consumed are of absolute fulfilment. Prior to implementation, development of the precision medicine is needed to be set as a wider scope's priority to enable more resources joining thus can be performed side by side by pertinent stakeholders as well as institutions to achieve powerful data which represent their identity and originality. We acknowledge that this is yet a beyond-the-sky concept. However, the assistance of integrated NGS and Big Data technologies, we are gradually advancing towards the direction.

References

[1] R. Guigo and M. de Hoon, “Recent advances in functional genome analysis,” *F1000Research*, vol. 7, p. F1000 Faculty Rev-1968, Dec. 2018, doi: 10.12688/f1000research.15274.1.

[2] S. Behjati and P. S. Tarpey, “What is next generation sequencing?” *Arch. Dis. Child. Educ. Pract. Ed.*, vol. 98, no. 6, pp. 236–238, Dec. 2013, doi: 10.1136/archdischild-2013-304340.

[3] K. Offit, “Personalized medicine: new genomics, old lessons,” *Hum. Genet.*, vol. 130, no. 1, pp. 3–14, Jul. 2011, doi: 10.1007/s00439-011-1028-3.

[4] G. S. Ginsburg and K. A. Phillips, “Precision Medicine: From Science To Value,” *Health Aff. (Millwood)*, vol. 37, no. 5, pp. 694–701, May 2018, doi: 10.1377/hlthaff.2017.1624.

[5] C. Gonzaga-Jauregui, J. R. Lupski, and R. A. Gibbs, “Human genome sequencing in health and disease,” *Annu. Rev. Med.*, vol. 63, pp. 35–61, 2012, doi: 10.1146/annurev-med-051010-162644.

[6] M. Prosperi, J. S. Min, J. Bian, and F. Modave, “Big data hurdles in precision medicine and precision public health,” *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, p. 139, Dec. 2018, doi: 10.1186/s12911-018-0719-2.

[7] T. Zhang et al., “BIGpre: A Quality Assessment Package for Next-Generation Sequencing Data,” *Genomics. Proteomics Bioinformatics*, vol. 9, no. 6, pp. 238–244, 2011, doi: https://doi.org/10.4258/annurev-med-051010-162644.

[8] H. Ayatollahi, S. F. Hosseini, and M. Hemmat, “Integrating Genetic Data into Electronic Health Records: Medical Geneticists’ Perspectives,” *Healthc. Inform. Res.*, vol. 25, no. 4, pp. 289–296, Oct. 2019, doi: 10.4258/healthcinform.2019.25.4.289.

[9] K. A. Phillips, J. R. Trosman, R. K. Kelley, M. J. Pletcher, M. P. Douglas, and C. B. Weldon, “Genomic sequencing: assessing the health care system, policy, and big-data implications,” *Health Aff. (Millwood)*, vol. 33, no. 7, pp. 1246–1253, Jul. 2014, doi: 10.1377/hlthaff.2014.0020.

[10] K. Y. He, D. Ge, and M. M. He, “Big Data Analytics for Genomic Medicine,” *Int. J. Mol. Sci.*, vol. 18, no. 2, p. 412, Feb. 2017, doi: 10.3390/ijms18020412.

[11] JASON, “The $100 Genome: Implications for the DoD,” *Notes*, pp. 1–58, 2010.

[12] U. Ferraro Petrillo, M. Sorella, G. Cattaneo, R. Giancarlo, and S. E. Rombo, “Analyzing big datasets of genomic sequences: fast and scalable collection of k-mer statistics,” *BMC Bioinformatics*, vol. 20, no. 4, p. 138, 2019, doi: 10.1186/s12859-019-2694-8.

[13] A. Belle, R. Thiagarajan, S. M. R. Soroshmehr, F. Navidi, D. Beard, and K. Najarian, “Review Article Big Data Analytics in Healthcare,” Aug. 2015, doi: 10.1155/2015/370194.

[14] S. Dash, S. K. Shakayawar, M. Sharma, and S. Kaushik, “Big data in healthcare: management, analysis and future prospects,” *J. Big Data*, vol. 6, no. 1, p. 54, 2019, doi: 10.1186/s40537-019-0217-0.

[15] J. K. Kulski, “Next-Generation Sequencing — An Overview of the History, Tools, and ‘Omic’ Applications,” *Intech Open*, pp. 1–59, 2015, doi: DOI: 10.5772/61964.

[16] S. T. Park and J. Kim, “Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing,” *Int. Neuropsychol. J.*, vol. 20, no. Suppl 2, pp. S76–S83, Nov. 2016, doi: 10.5213/inj.1632742.371.

[17] L. Liu et al., “Comparison of next-generation sequencing systems,” *J. Biomed. &amp;
[18] C. Di Resta, S. Galbiati, P. Carrera, and M. Ferrari, “Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities,” *EJIFCC*, vol. 29, no. 1, pp. 4–14, Apr. 2018.

[19] J. Baselga *et al.*, “AACR Cancer Progress Report 2015,” *Clin. Cancer Res.*, vol. 21, no. 19 Suppl, pp. S1–S128, Oct. 2015, doi: 10.1185/1078-0432.CCR-15-1846.

[20] M. Lynch, “Mutation and Human Exceptionalism: Our Future Genetic Load,” *Genetics*, vol. 202, no. 3, pp. 869–875, Mar. 2016, doi: 10.1534/genetics.115.180471.

[21] P. Park *et al.*, “Next-Generation Sequencing-Based Cancer Panel Data Conversion Using International Standards to Implement a Clinical Next-Generation Sequencing Research System: Single-Institution Study,” *JMIR Med. informatics*, vol. 8, no. 4, pp. e14710–e14710, Apr. 2020, doi: 10.2196/14710.

[22] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, “Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration,” *Brief. Bioinform.*, vol. 14, no. 2, pp. 178–192, Apr. 2012, doi: 10.1093/bib/bbs017.

[23] P. Kaur, M. Sharma, and M. Mittal, “Big Data and Machine Learning Based Secure Healthcare Framework,” *Procedia Comput. Sci.*, vol. 132, pp. 1049–1059, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.020.

[24] Y. Zhang, M. Qiu, C. W. Tsai, M. M. Hassan, and A. Alamri, “Health-CPS: Healthcare cyber-physical system assisted by cloud and big data,” *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, 2017, doi: 10.1109/JSYST.2015.2460747.

[25] R. R. Gullapalli, M. Lyons-Weiler, P. Petrosko, R. Dhir, M. J. Becich, and W. A. LaFramboise, “Clinical integration of next-generation sequencing technology,” *Clin. Lab. Med.*, vol. 32, no. 4, pp. 585–599, Dec. 2012, doi: 10.1016/j.cll.2012.07.005.

[26] J. Shendure and H. Ji, “Next-generation DNA sequencing,” *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1135–1145, 2008, doi: 10.1038/nbt1486.

[27] J. L. Jameson and D. L. Longo, “Precision Medicine — Personalized, Problematic, and Promising,” *N. Engl. J. Med.*, vol. 372, no. 23, pp. 2229–2234, May 2015, doi: 10.1056/NEJMsb1503104.

[28] B. Ristevski and M. Chen, “Big Data Analytics in Medicine and Healthcare,” *J. Integr. Bioinform.*, vol. 15, no. 3, p. 20170030, 2018, doi: https://doi.org/10.1515/jib-2017-0030.

[29] A. O’Driscoll, J. Daugelate, and R. D. Sleator, “‘Big data’, Hadoop and cloud computing in genomics,” *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–781, 2013, doi: https://doi.org/10.1016/j.jbi.2013.07.001.