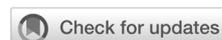


Rapid advancement in cancer genomic big data in the pursuit of precision oncology

Tiara Bunga Mayang Permata, Sri Mutya Sekarutami, Endang Nuryadi, Angela Giselvania, Soehartati Gondhowiardjo



pISSN: 0853-1773 • eISSN: 2252-8083
<https://doi.org/10.13181/mji.rev.204250>
Med J Indones. 2021;30:81–5

Received: October 08, 2019
Accepted: September 01, 2020
Published online: January 13, 2021

Authors' affiliations:
 Department of Radiation Oncology,
 Faculty of Medicine, Universitas
 Indonesia, Cipto Mangunkusumo
 Hospital, Jakarta, Indonesia

Corresponding author:
 Tiara Bunga Mayang Permata
 Department of Radiation Oncology,
 Faculty of Medicine, Universitas
 Indonesia, Cipto Mangunkusumo
 Hospital, Jalan Pangeran Diponegoro No.
 71, Kenari, Senen, Central Jakarta 10430,
 DKI Jakarta, Indonesia
 Tel/Fax: +62-21-3921155
E-mail: dr.mayangpermata@gmail.com

ABSTRACT

In the current big data era, massive genomic cancer data are available for open access from anywhere in the world. They are obtained from popular platforms, such as The Cancer Genome Atlas, which provides genetic information from clinical samples, and Cancer Cell Line Encyclopedia, which offers genomic data of cancer cell lines. For convenient analysis, user-friendly tools, such as the Tumor Immune Estimation Resource (TIMER), which can be used to analyze tumor-infiltrating immune cells comprehensively, are also emerging. In clinical practice, clinical sequencing has been recommended for patients with cancer in many countries. Despite its many challenges, it enables the application of precision medicine, especially in medical oncology. In this review, several efforts devoted to accomplishing precision oncology and applying big data for use in Indonesia are discussed. Utilizing open access genomic data in writing research articles is also described.

KEYWORDS cancer genetic database, oncology, personalized medicine

The medical world is overwhelmed by the supply and rapid advancement of massive data, which are commonly referred to as big data. Open access genomic databases, along with analytic platforms, are available worldwide to ease the challenges in utilizing such complicated data. For research advancement, this information, along with tools, is invaluable. Most articles published in high-impact factor journals allot a special portion for biostatistical analysis involving big data to improve their paper quality. Therefore, the scientific world has recognized the importance of genomic big data.

In clinical settings, genetic sequencing development has facilitated a routine clinical sequencing for patients with cancer. The sequencing helps to determine the specific gene mutations in individual and guide physicians in providing an appropriate therapy. This principle of precision medicine has been applied

to oncology. In this review, the history of genomic big data, the efforts devoted to precision oncology, specifically in the development of genomic databases globally and locally, and the challenges encountered in this field are discussed.

Big data

Big data is considered as the "new oil" in today's industries, represented by the presence of companies such as Google and Facebook in the top list of global companies.¹ The term big data itself is defined as massive volumes of complex and interconnectable information. Many innovations in medical technology, such as sequencing techniques, breakthroughs in molecular biology, biomedical informatics, and computing, have been launched; consequently, the field of big data has been rapidly developed in medicine.² The implication of big data

in medicine is not limited to genomics or other omic data; it encompasses medical (from electronic medical records), environmental, financial, geographical, and social media information. The size of data, especially genomic data, is rapidly growing, so this type of data is estimated to be the single largest data contributor to the world and will likely exceed the data size of YouTube or even astronomy, the current largest data contributor.³

Open access genomic data portals

The Cancer Genome Atlas

Technological advancements in molecular genomics have provided insights into genetic events, which clinically determine the phenotype of a patient with cancer. However, many complex molecular events or pathways have yet to be elucidated or revealed. Therefore, in 2005, the National Institute of Health of the USA launched The Cancer Genome Atlas (TCGA), a colossal project, to compile profiles and analyze human tumor samples in large volumes at DNA, RNA, protein, and epigenetic levels.^{4,5} After 10 years of its launch, the government-funded TCGA has coordinated data collection and genomic investigation of 33 cancer types and has consequently made many vital discoveries in cancer based on individual genetic variations and their surrounding pathways, which can be accessed at <https://cancergenome.nih.gov/publications>.⁶⁻⁸ The TCGA team has successfully collected genetic information from 11,160 patients with the collaboration of 20 different institutions in the USA and Canada.^{6,9}

The Cancer Cell Line Encyclopedia

Human cancer cell lines are still the primary modality in cancer biology research and drug discovery research because of the infinite possibilities of manipulations in experiments. Consequently, understanding the genetic characteristics of cell lines used is necessary to strengthen the analysis of *in vitro* findings, which can be translated into hypotheses, research method formulation, and planning. The Cancer Cell Line Encyclopedia (CCLE) project was established to address this need and initiate large-scale data collection for 947 human cancer cell lines, which cover 36 different cancer types. The genetic characteristics of cell lines have been sequenced and displayed (more than 1,600 genes), and general recurrent mutations in different cancer types have been revealed, i.e., 492 mutations in 33 identified oncogenes and tumor suppressor genes.¹⁰ CCLE is also

an open access portal at <https://portals.broadinstitute.org/ccle>, which currently provides information for 1,457 cell lines.

Growing presence of genomic data analysis in scientific publications

The availability of big data has opened many opportunities. Individuals in science or medicine are interested in obtaining and applying genomic data to clinical settings. Researchers have also shifted to utilizing big data in a high-level scientific publication. In current scientific publications, even high-impact factor or Q1 journals, original articles have described biostatistical analysis by using the TCGA database or other similar portals to support their hypothesis and *in vitro* or clinical findings.

Bioinformatic analysis was applied in our previous studies to explore the role of base excision repair (BER) in the Programmed death-ligand 1 (PD-L1) expression in cancer cells.¹¹ The TCGA database was utilized to investigate the mutation status of BER genes and the mRNA expression levels of PD-L1. Neoantigen levels were obtained from another portal, The Cancer Immunome Atlas, accessible at <https://tcia.at/home>, analyzed, and published in high-level international journals.^{11,12} Exploring genomic big data in scientific works has been proven useful in strengthening the quality and quantity of academic findings in published articles and our own experiences.

Other examples were abundant and interesting to study different ways to apply big data analysis in scientific publications; among them are studies on *ARID* genes.¹²⁻¹⁵ Pan et al¹³ described the role of a chromatin regulator in the resistance of tumors to T cell-mediated cytotoxicity in the journal *Science* in 2018. They showed the correlation of the expression of several genes, including *ARID2*, *PBRM1*, and *GZMB*.¹³ In *Nature Medicine* in 2018, Shen et al¹⁴ analyzed 26 cancer types in the TCGA samples to show the difference in the mutation load in tumors with or without mutation in the *ARID1A* gene (a part of the chromatin remodelling complex).

Genomic big data is considered not only as complementary data but also as a sole modality for studies, such as those successfully published in *Nature Communications*. In 2017, Cortes-Ciriano et al¹⁶ described the microsatellite instability profile in different types of cancer by using the TCGA data. They analyzed 8,000 exomes and 1,000 whole genomes from 23 different cancer types and proposed a sequencing

panel to check the microsatellite instability (MSI) levels in patients who are considered for immunotherapy. Therefore, clinicians and scientists agreed that biostatistical analysis has potential for application in the improvement of studies or scientific paper writing. Big data has been widely used by the international scientific society in high-quality publications.

Challenges and tools for genomic data analysis

However, utilizing or understanding genomic data is complicated. For people without a background in biostatistics or computing, even downloading data into generally used computer software can be a major challenge.¹⁷ In addition to this technical challenge, many genomic aberrations and a combination of important aberrations, which drive carcinogenesis and bystander mutations that have no contributions to phenotypes, are found in cancer genomic data.¹⁸ Therefore, a unique skill is needed to process complicated data and understand medical genetics so that gene sequencing results can be sorted in a logical thinking framework, i.e., which data should be considered for analysis and which ones should be disregarded.¹⁹ This task should be explored in the near future, considering that human resources who have background in computing and biology are scarce globally.¹⁸

Fortunately, the number of sites or portals providing data analysis, including internal analysis projects from various databases, such as the TCGA and CCLE, and external portals, such as the TCGA's Clinical Explorer or cBioPortal, is increasing.¹⁷ The TCGA team has its own analysis project, namely, the Pan-cancer analysis (PCA). The PCA was founded in 2013 to study the first 12 tumor types profiled by the TCGA.⁴ This project has been continually improved in terms of accuracy and coverage; the discovery on association between high-levels of PD-1/PD-L1 in more than 300 MSI-high tumors has also been enhanced.⁷

The Tumor Immune Estimation Resource (TIMER), another external analysis portal, can be freely accessed at <https://cistrome.shinyapps.io/timer/>; it also has a straightforward video tutorial on their homepage.²⁰ In TIMER, TCGA samples are used, and the infiltration of immune components, including B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells, in tumors is analyzed. Many features are user friendly; one of the most important features is the determination of the correlation between any two gene expression levels, which can be adjusted in terms

of tumor purity.²⁰ Analysis involving TIMER has been published in several well-known publications, so it may also be utilized in our studies.^{11,13,21}

Genomic data utilization in the region

Genomic data have been used not only in Western countries but also in Asia. In the International Cancer Genome Consortium list of contributing institutions, 15 out of 99 institutions are from Asia, showing the potential of genomic data for application in this region. The five Asian countries that have made significant contributions are China, Japan, South Korea, Singapore, and India.²²

As reported in the Federations of Asian Radiation Oncology 2019 Meeting, Pakistan, Thailand, and Bangladesh are planning to launch their genomic projects. Malaysia also is already running active research projects. Most Asian countries have agreed that such projects must be pursued sooner rather than later. However, the capabilities of each country vary and have unique challenges and obstacles.²³

In Indonesia, many scientists have been familiarized with big data. However, few of them are dedicated to studying cancer genomic big data, which are unique. Publications under the name of this country have gradually emerged in international collaborations.^{11,12,15,24} Nevertheless, cancer genomic data may be considered in academic papers if and only if we decide to study and utilize them.

Future prospects of precision oncology

Precision medicine is defined as a concept in which health care is designed individually and matched with a patient's genetic information, lifestyle, and environment.²⁵ In the current big data era, personalized medicine is achieved.¹⁸ Scientific advancement in understanding genetic mutation or amplifications has shown a significant impact on cancer treatment. A major progress can be observed in choosing a group of patients to be included in selected clinical trials, in restricting drug indications to patients who have no response to some treatments, and in estimating the toxicity risk of certain therapies.¹⁹

In applying precision medicine to oncology, the goal in cancer control is to cover aspects of prevention, early detection, and treatment. However, a comprehensive understanding of cancer as the accumulation of phenotypic consequences of genetic, genomic, and epigenetic changes in cancer cells is

needed to achieve this goal; furthermore, numerous interactions occur in tumor microenvironments. With such understanding, strategies that can be applied to accomplish this goal can be planned to develop the most suitable therapies for specific patients' subpopulations.¹⁸

In the USA, researchers from the Memorial Sloan Kettering (MSK) Cancer Center have applied the test of the Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) to detect changes in cancer-related genes in tumors and matching normal periphery blood samples. The panel included 341 genes in the first development of this test. Since then, this number has been growing. In a 2017 publication, 37% of 10,000 tested patients had at least 1 actionable mutation, and 11% of the first 5,009 patients were enrolled in clinical trials based on their genes. Thus far, the MSK-IMPACT test has covered 468 genes and has been approved by the Food and Drug Administration.²⁶

Other developed countries have conducted related studies. In 2018, Japan has started sequencing via the MSK-IMPACT panel by collaborating with the MSK Cancer Center. Tumor and matching normal specimens

from three teaching hospitals in Japan were sent to the MSK. Based on sequencing results, patients with actionable gene mutations were enrolled in clinical trials or given appropriate targeted therapy in Japan.²⁷ Consistent with this process, the National Cancer Center in Japan also developed a sequencing panel, which was launched as the Cancer Genome Screening Project for Individualized Medicine in Japan (SCRUM-Japan) project. This project aimed to identify the changes in Japanese patients' oncogenes and facilitate sample recruitments for clinical trials with targeted therapies.²⁷ With Japan's rigorous efforts and multisectoral support, the cancer gene panel sequencing can be included in the national health insurance.²⁶

A scientific discovery in cancer genetics should be comprehensively investigated before it can be included in cancer therapy and applied to clinical settings. For instance, the discovery of the *KRAS* gene mutation took 30 years until its mutation status could be utilized in cancer treatment; instead of being a drug target, it was finally used as a marker of tumor resistance to anti-epidermal growth factor receptor (EGFR) drugs. Nowadays, patients with colon or lung cancer are recommended to have their *KRAS* mutation status checked before they undergo EGFR-targeted therapy.¹⁸

Although such initiatives take time and require coordinated efforts, this rapidly advancing field should be introduced to our daily critical thinking and integrated into our research and publications (Figure 1). As scientific publication expands exponentially, its translation to clinical applications accelerates. When all stakeholders, including pharmaceutical staff, clinicians, researchers, insurance company personnel, and governments, apply the rapid technological advancement in leading the growth of cancer translational studies, precision oncology may be achieved.

In conclusion, multidisciplinary efforts, including basic science, clinical trials, and therapeutic application to patients, should be directed at a multinational level to visualize the ultimate goal of translating discoveries at genomic levels into evidence-based clinical practice.

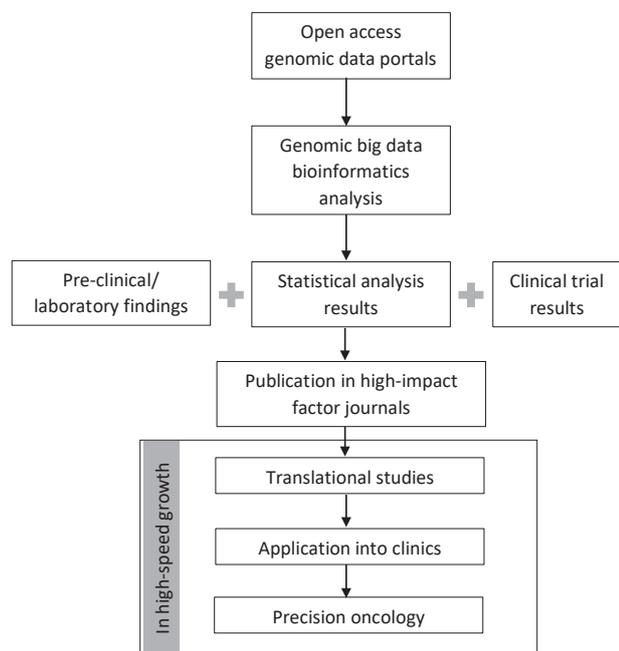


Figure 1. Utilization of genomic big data analysis in the pursuit of precision oncology. The widely accessible portals of genomic databases are now available for use in bioinformatic analysis in any type of oncology research. The result of such *in silico* analysis can be included in a scientific paper to achieve a comprehensive presentation and obtain results from *in vitro* experiments and clinical trials or observations. As these data make their way into scientific publications, they drive the rapid visualization of precision oncology

Conflict of Interest

The authors affirm no conflict of interest in this study.

Acknowledgment

None.

Funding Sources

None.

REFERENCES

1. The Economist. The world's most valuable resource is no longer oil, but data [Internet]. The Economist; 2017 [cited 2018 Aug 25]. Available from: <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>.
2. Hinkson IV, Davidsen TM, Klemm JD, Chandramouliswaran I, Kerlavage AR, Kibbe WA. A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Front Cell Dev Biol.* 2017;5:108.
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomic? *PLoS Biol.* 2015;13(7):e1002195.
4. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013;45(10):1113–20.
5. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol.* 2015;19(1A):A68–77.
6. National Cancer Institute. The Cancer Genome Atlas: Program Overview [Internet]. 2016 [cited 2018 Aug 20]. Available from: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>.
7. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell.* 2018;173(2):371–85.e18.
8. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell.* 2018;173(2):321–37.e10.
9. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. *Cell.* 2018;173(2):400–16.e11.
10. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* 2012;483(7391):603–7.
11. Permata TB, Hagiwara Y, Sato H, Yasuhara T, Oike T, Gondhowiardjo S, et al. Base excision repair regulates PD-L1 expression in cancer cells. *Oncogene.* 2019;38:4452–66.
12. Sato H, Niimi A, Yasuhara T, Permata TB, Hagiwara Y, Isono M, et al. DNA double-strand break repair pathway regulates PD-L1 expression in cancer cells. *Nat Commun.* 2017;8:1751.
13. Pan D, Kobayashi A, Jiang P, de Andrade LF, Tay RE, Luoma AM, et al. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. *Science.* 2018;359(6377):770–5.
14. Shen J, Ju Z, Zhao W, Wang L, Peng Y, Ge Z, et al. ARID1A deficiency promotes mutability and potentiates therapeutic antitumor immunity unleashed by immune checkpoint blockade. *Nat Med.* 2018;24(5):556–62.
15. Nuryadi E, Sasaki Y, Hagiwara Y, Permata TB, Sato H, Komatsu S, et al. Mutational analysis of uterine cervical cancer that survived multiple rounds of radiotherapy. *Oncotarget.* 2018;9(66):32642–52.
16. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun.* 2017;8:15180.
17. Lee HJ, Palm J, Grimes SM, Ji HP. The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations. *Genome Med.* 2015;7:112.
18. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med.* 2011;17(3):297–303.
19. Dancy JE, Bedard PL, Onetto N, Hudson TJ. The genetic basis for cancer treatment decisions. *Cell.* 2012;148(3):409–20.
20. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 2017;77(21):e108–10.
21. Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 2016;17(1):174.
22. Zhang J, Bajari R, Andri D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium data portal. *Nat Biotech.* 2019;37:367–9.
23. Permata TB, Utami IG, Gondhowiardjo S. Towards precision oncology and the need for Asian cancer genomic big data. In: FARO 3rd Annual Meeting: Cancer free world, making it a reality. Shenzhen: Federations of Asian Radiation Oncology (FARO); 2019.
24. Nuryadi E, Permata TB, Komatsu S, Oike T, Nakano T. Inter-assay precision of clonogenic assays for radiosensitivity in cancer cell line A549. *Oncotarget.* 2018;9(17):13706–12.
25. Hodson R. Precision medicine. *Nature.* 2016;537:S49.
26. Kohno T. Implementation of “clinical sequencing” in cancer genome medicine in Japan. *Cancer Sci.* 2018;109(3):507–12.
27. Saito M, Momma T, Kono K. Targeted therapy according to next generation sequencing - based panel sequencing. *Fukushima J Med Sci.* 2018;64(1)9–14.