Artificial Intelligence Multiomics in Precision Oncology

Artificial Intelligence Multiomics in Precision Oncology

By

Ruby Srivastava

Cambridge Scholars Publishing

Artificial Intelligence Multiomics in Precision Oncology

By Ruby Srivastava

This book first published 2023

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-9520-2
ISBN (13): 978-1-4438-9520-0

Dedicated to my parents

(Late) Dr. Mohan Swaroop Srivastava

Mrs. Manju Srivastava

With heartfelt gratitude to those who were my strongest support in the toughest times……..

TABLE OF CONTENTS

PREFACE

Cancer is the second most common disease-related cause of death among patients below the age of 70 years and is considered the most significant cause of death in the 21st century. It is the major obstacle to the increase of global life expectancy. It has spread in 91 countries and the fast growth of cancer incidence and mortality now attracts significant attention from the governments, the medical industry and the scientific community. In answer to the question of how we can reduce the cancer-related death rate, there have been increasing advances in the development of effective and safe drugs for the treatment of cancers. In the recent years, Artificial Intelligence (AI) has shown immense potential in the field of health sciences and has contributed significantly at each stage of cancer. The contributions of AI include reliable early detection, stratification, determination of infiltrative tumor margins during surgical treatment, drug and therapy response, tracking tumor progression and resistance to treatments, prediction of tumor aggression, evolution pattern of tumors, and its repetitiveness. Machine learning (ML) and AI have enabled the development of new ways to analyze big datasets more quickly and cost-effectively. The deep learning mechanisms have solved problems and enhanced the decision-making abilities for 'multiomics' data (genomics, epi-genomics, transcriptomics, proteomics, and metabolomics), and 'non-omics' data (medical/mass-spectrometry imaging, patient clinical history, treatments, and disease endemicity), helping to overcome the challenges of accurate detection, prediction, diagnosis, and treatment of cancer patients. As AI reduces the fuzziness and randomness of data handling, it can serve as the primary choice for data mining and big data analysis. The role of AI in handling the big data analysis includes: (a) analysis of complex and heterogeneous multiomics and/or interomics datasets; (b) providing holistic disease molecular mechanisms; (c) identification of diagnostic and prognostic markers; and (d) monitoring a patient's response to drugs and therapy, and their recovery. Precision medicine is used to treat cancer patients given the heterogeneity of diseases. Oncology has seen extensive benefits from AI in the treatment of cancers including early detection, prediction, and treatment for future outcomes. Next generation sequencing (NGS) has addressed certain demands and has become a prominent player in the revolution of precision oncology. The various clinical applications of NGS include risk prediction, early detection of disease, diagnosis by sequencing

and medical imaging, accurate prognosis, biomarker identification, and the identification of therapeutic targets for novel drug discovery. The large datasets generated by NGS require expertise in the bioinformatic resources to analyze clinically relevant and significant data. The application of AI is crucial in aspects of radiology such as X-rays, ultrasounds, computed tomography (CT/CAT), magnetic resonance imaging (MRI), positron-emission tomography (PET), and digital pathology. Highly specialized algorithms are used to analyze the data with high accuracy and speed. The datasets generated include information about variants with classifications such as benign, likely benign, variant of unknown significance, likely pathogenic, and pathogenic. Classifying and categorizing these data can be useful for the prognosis and diagnosis of cancers. The automation of healthcare systems is particularly important in resource-deprived developing countries. Shortages of well-trained healthcare workers and specialists are a major concern which can be addressed by implementing AI systems that can diagnose diseases more quickly. The other advantage of AI is that it can reduce the burden of health records and eliminate mandatory administrative formalities. Automated AI-enabled systems will allow physicians to sort and analyze patient health records, so that doctors can take clinical decisions. AI will help to ease the well-documented workload of physicians. Moreover, if the clinical data of cancer patients became globally accessible, it could be used by doctors all over the world to reliably predict the future risk of diseases.

However, the implementation of AI in the health sector still faces many barriers despite its obvious usefulness. With computational automation, there is a flood of big data and costs. As these data depend on the specialized requirements of the fast processing of data, AI systems have become very expensive. These systems also require additional quality processes. If the data can be predicted and interpreted properly, with expertise and with an understanding of the system, AI systems can offer accurate data and image analysis. A major hurdle to the advancement of AI is found in the ethical challenges which occur in the healthcare industry. Ethical guidelines are required to protect the patient's safety and privacy. Public access to these data will have legal consequences. With advances in AI technology, AI will definitely overcome these limitations and challenges. Also, other molecular characterization technologies such as multiomic approaches, drug/therapy treatments, and combination treatments of drugs beyond monotherapy will increase the clinical utility and scope of personalized treatments for cancer patients in the future.

Welcome to the first edition of the book, Artificial Intelligence Multiomics in Precision Oncology. The aim of the book is to introduce to readers and researchers in various fields the latest development in big data multiomics techniques and its advancement with AI in precision oncology. As cancer is a heterogeneous disease, AI approaches should enable non-effective treatments to be quickly replaced by alternative combination therapies so that patients can recover more quickly and survive longer. Despite many challenges and ethical issues, we have 'long-term optimism' about the role AI Multiomics can play in precision treatments for cancer patients. Hopefully, AI based tools will also develop to guide the unexplored personalized follow-up treatments for cancer patients in the future.

I am very thankful to Helen, book publishing manager for Cambridge Scholars Publishing for her unconditional support at every level. *Thank you so much Helen*. Last but not least, I would like to thank my husband Amit, daughter Arghyaa, son Aryan, and all my family for their continuous support and encouragement at every level in the writing of this book.

ACKNOWLEDGEMENTS

I acknowledge the financial assistance of the DST WOSA project (SR/WOS-A/CS-69/2018). I am also thankful to my Mentor, Dr. Shrish Tiwari, Bioinformatics, Centre for Cellular and Molecular Biology-CSIR, and Dr. G. Narahari Sastry, Director, CSIR-NEIST for their support.

I am thankful to the proof reader of this book, Mr. John Ingamells. His patience and unconditional support have been my utter strength to finish the book. *Thank you so much John!*

CHAPTER 1

INTRODUCTION

Cancer is a group of diseases which is characterized by the uncontrolled growth and spread of abnormal cells. It can result in death if not treated properly. Although the causes of cancer development are not completely understood, numerous factors are known to increase risk, including many that are potentially modifiable (e.g., tobacco use and excess body weight) and others that are not (e.g., inherited genetic mutations). These risk factors may act simultaneously or in sequence to initiate and/or promote cancer growth. More than 1.9 million new cancer cases are expected to be diagnosed in the United States in 2022; this excludes basal cell and squamous cell skin cancers and carcinoma *in situ* (non-invasive cancer) except for urinary bladder. The cancer immunity cycle with the different roles of cells is shown in Figure 1.

Advances in multidimensional omics technologies from Next Generation Sequencing (NGS) to Mass Spectrometry (MS) have created a pool of useful information which can be used in the treatment of diseases. AI mediated data integration enables the understanding of complex disease systems by describing all the biomolecular entities from DNA to metabolics. Multiomics approaches have diversified applications, not only in oncology, but also in veterinary medicine (1), microbiology (2), agriculture sciences (3), biofuel (4), biomedical sciences (5,6), and many others. The speed, accuracy, and affordability of NGS data have helped the activation of personalized treatment based on the disease, driving molecular alterations in cancer treatment. NGS data has been tested in many healthcare settings and it is in advanced use in oncology. Physicians match the sequencing data of a patient's tumor with the designed therapies to target the genetic alterations driving the tumor growth. These therapies are known as sequencing-matched therapies. The entire diagnosis depends on the efficient use of genomic data, clinical information, and patient preferences in making clinical decisions that improve outcomes by matching each patient with the therapy best suited to treat their cancer. The incorporation of NGS into treatment improves patient outcomes in

both treatment response and disease-free survival. It does, however, generate controversy related to the insurance coverage of the patients and this topic is still under debate. NGS technologies have generated a significant amount of information related to the disease progression, the understanding of cancer biology, and treatments for cancer patients.

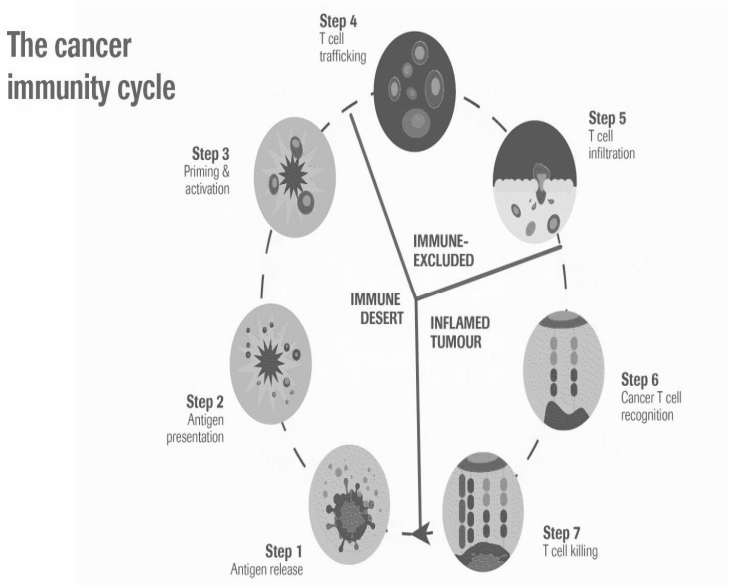

Figure 1: The cancer immunity cycle showing the different roles of cells. The cycle starts with the release of antigens and ends with the formation of immunity to the cancer cells. Reused from https://twitter.com/roche/status/1097496056072495106?lang=ca).

The term ‘Transcriptomics’ represents active genes as well as long noncoding RNAs, short RNAs such as microRNAs, and small nuclear RNAs in a defined physiological condition. The two main applications of research include transcript discovery and RNA quantification. Transcriptomic analysis evaluates overall transcripts in a metabolic process, while the targeted approach provides information regarding known genes. A differential expression (DE) of protein-coding RNA provides insight into the disease mechanism, integrated with genomics and proteomics to discover novel genes and their functional relevance. Noncoding RNAs have regulatory functions in several metabolic diseases, neurological disorders, and cancers. Transcriptome is directly correlated to

any epigenomic change that manifests cancer. The integration of transcriptomics and epigenomic data could thus extend our understanding of cancer biology in various types of cancers.

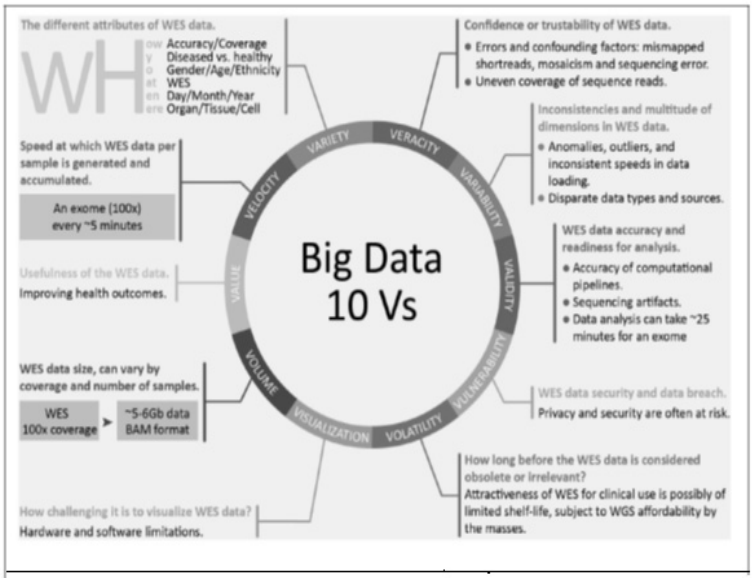

Figure 2: The schematic representation of 10Vs (big data) for whole-exome sequencing. (Reused from Suwinski P. et al. Front. Genet. 2019;10:49.)

Genomic data relies entirely on the nucleotide sequences, including expressed sequence tags (ESTs), cDNAs, and gene arrangements on the respective chromosomes. Rapid advances in NGS data and *in silico* approaches lead to high throughput data for whole-genome sequencing and epigenomics. WGS explores all types of genomic alterations in cancer and provides information on the range of driver mutations and mutational signatures for the non-coding regions in cancer genomes. WGS is thus a powerful tool to understand cancer genomics that contains unpredictable numbers of point mutations, fusions, and other aberrations. Though the target approaches using whole-exome sequencing (WES) are much easier to analyse, little information related to untranslated, intronic, or intergenic regions is missed out, which also affects the molecular pathogenesis of cancer [7]. The 10 Vs (volume, velocity, variability, variety, veracity, validity, vulnerability, volatility, visualization, and value) for whole-exome sequencing are given in Figure 2.

There are several limitations associated with the WGS data. Many of the clinical reports lack a comprehensive clinical annotation linking genomic events to specific cancer types, diagnosis, and treatment of responses. Also, most of the genomic data is focussed on the target approaches. Since most of the preliminary studies are performed on the untreated cancers, it does not provide insight into the response for the treatments [8]. It is, therefore, better to integrate the cancer genomic data with clinical physiology data for treatment efficacy. All the multiomics techniques related to the system biology (transcriptomics, epigenomics, genomics, proteomics, and metabolics) (Figure 3) will be discussed in detail in the next few chapters.

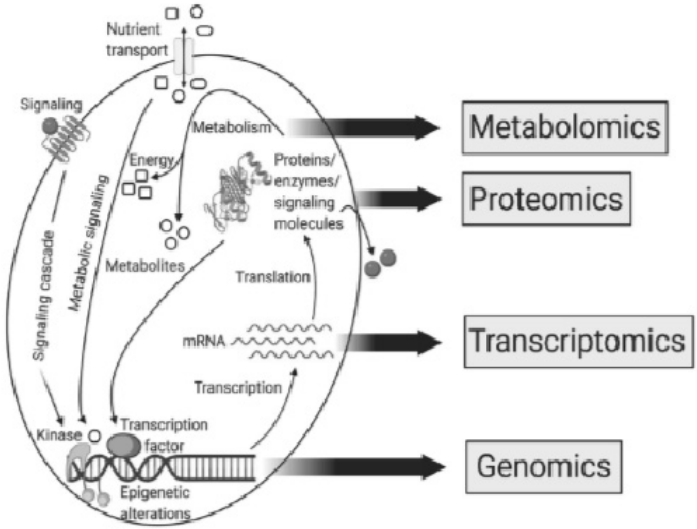

Figure 3: Representation of multiomics approaches between system biology. (Reused from Schmidt. DR. et al. CA a Cancer Journal of Clinicians 2021; 71(2): 333–358.)

According to Prof. Marc Wilkins, a proteome is the entire complement of proteins that is or can be expressed by a cell, tissue, or organism at a given time. Proteomics reveal information about cellular/molecular responses to (epi-) genomics, environmental alterations, and their feedback responses. Mass spectrometry (MS) based proteomic approaches include explorative proteomics, targeted proteomics and MS imaging, which focus on all proteins of a cell/organism, subsets of proteins, and location of proteins. The explorative workflow of proteomics includes sample preparation,

mass spectrometry, and data analysis. MS also has applications beyond disease diagnostics. MS can monitor the feedback responses towards therapy, identification of drug toxicity, and discovering new biomarkers. For this purpose, high quality datasets are required. Other requirements include improvements in MS-instrument quality and robustness, automated sample processing, robust data analysis pipelines, and online automation (cloud computing) to integrate results, datasets, and data portability. Furthermore, the Human Proteome Organization (HUPO) has set up guidelines for sample collection viz. selecting appropriate disease controls, categorizing disease and sub-disease status [9], storage to rule-out pre-analytical variables (including patient and instrumental factors) that contribute to a large extent of variation, calibrating MS instrument for data-quality assurance, data reporting for untargeted and targeted [10] analysis. An amalgamation of proteomics data with interomics data and cancer histopathological images using AI has advanced the identification of metabolic pathways. Post-translation modifications, including phosphorylation, glycosylation, ubiquitination, and nitrosylation enrich the protein repertoire (protein isoforms) and affects protein functions (transport, enzymatic activity), and intracellular signaling pathways in cancer. The classification of specific reforms provides unmatched clinical sensitivity and specificity. Metabolomics is a newly emerging 'omics' field which is used for comprehensive and simultaneous systematic determination of metabolite levels in the metabolome and their changes over time as a consequence of stimuli. Metabolomes are dynamic and complete sets of small-molecule metabolites, while metabolites are the intermediates and products of metabolism. There are primary and secondary metabolites. Metabolites have multiple categories such as antibiotics, pigments, carbohydrates, fatty acids, and amino acids. Metabolomics is the systematic analysis of small molecules (<1kD) within cells, biofluids, tissues, or organisms involved in primary or secondary metabolic processes. Metabolite changes significantly during the process of normal growth and development and/or exposure to stress, allergens, and disease conditions [11-13], which relates strongly to the final clinical phenotype. Metabolics thus enhance the molecular understanding of disease mechanisms, progression, response to drugs/treatments, and recurrence probability. The workflow of metabolics comprises metabolite extractions, separation by liquid/gas chromatography, capillary electrophoresis and ion mobility, detection by mass spectrometry (MS), or nuclear magnetic resonance (NMR) spectroscopy and data analysis. In recent years, there have been advances in the applications of metabolics as a result of the discovery and development of soft ionization tools such as

electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). Several separation-free MS techniques are available, including direct infusion-MS, MALDI-MS, mass spectrometry imaging (MSI), and direct analysis in real-time mass spectrometry. High-throughput MSI analysis is a powerful tool for the identification of biomarker, tracking drugs and its metabolites, imaging drug-response at cellular-level. These tools also identify unique and specific biomarkers (lipid signature) and therapeutic targets to classify various types of cancer. MS-based metabolomics revealed four metabolites (oleanoic acid, taurochenodeoxycholate, palmitic acid, and d-sphingosine) as highly discriminative potential prognostic markers.

By definition a biomarker is ‘a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention’. Biomarkers are categorized as diagnostic biomarkers, prognostic biomarkers, predictive biomarkers, and predisposition biomarkers. Diagnostic biomarkers are used to determine the specific health disorder of the patient; prognostic biomarkers help to chart the likely course of the disease; predictive biomarkers indicate the probable response to a particular drug; and predisposition biomarkers indicate the risk of developing a disease. Biomarkers can positively impact the treatment of patients by predicting individual disease risk, allowing early detection of disease, which often increases the effectiveness of treatment. They also improve diagnostic classification which in turn may promote personalized treatment. They also enable monitoring of the progress of a given therapy. Till now very few omics-derived biomarkers have made it into clinics. The complexity of cellular processes involved in tumor formation, heterogeneity of neoplasia (different tumors, intertumoral, and intratumoral), non-optimal study design, and poor methodological robustness and reproducibility are the main factors for the large gap between the number of omics-based biomarkers found in research and those introduced in clinics [14]

The quality, quantity, and availability of tumor tissue from cancer patients pose challenges to the clinical implementation of precision medicine. The processing of formalin-fixed, paraffin-embedded fragments can alter nucleic acids, and low tumor content in tumor samples can decrease the sensitivity of tests and lead to false-positive mutation calls. The biopsies collected at a single time point may not account for intratumoral heterogeneity in space or time. To overcome these limitations, multiple biopsies are needed which is inhibited by the extent of resources needed and by the requirement to ensure patient safety. There are circulating

tumor-specific markers which include circulating tumor cells (CTC) or circulating tumor DNA (ctDNA), as well as RNAs, proteins, or metabolites that are present in body fluids such as blood, urine, and peritoneal or cerebrospinal fluid [15]. Ideally, biomarkers should be analyzed in non-invasive biospecimens like blood, plasma, serum, urine, saliva, or stool. Liquid biopsies are easily accessible through minimally invasive procedures that can be repeated to provide a dynamic and longitudinal assessment of tumor-specific diagnostic, prognostic, or predictive biomarkers. Yet the Omics-derived biomarkers are still not helping pediatric oncologists in decision-making. Drug target discovery is a crucial step for the development of cancer drugs and precision therapeutics. Traditional drug target discovery consists of biomolecules with a confirmed mechanism of action, which is selected in a series of studies [16, 17]. Over the last decade, putative drug targets have been identified through the latest NGS approaches in combination with experimental validation. This has included overexpression or knockdown by RNAi and the use of transgenic animals and model organisms [18]. Multiomics approaches may allow systematic assessment of drug discovery for personalized cancer therapy and improve the efficacy of chemotherapy [19, 20]. Refining molecular-defined subsets of patients can provide information on drug response and resistance among the individual patients. In recent studies the expression of lncRNA, miRNA, mRNA, methylation, and the profile of somatic mutations with the expression of drug response-related lncRNAs were integrated. Fourteen cancer subtypes from TCGA multiomics datasets were analyzed, revealing 40 driver genes associated with the Wnt, Notch, Hedgehog, JAK/STAT, NK-KB, and MAPK signaling pathways [21]. Among them, well-known driver genes such as EGFR, ERBB2, PIK3CA, and KRAS were confirmed to be upregulated in several cancers, and DCUN1D1 and NSD3 were identified as new driver genes. The success of trastuzumab (an agent targeting HER2) in breast cancer has opened a new era of novel druggable targets in cancers. Proteomic analysis of 105 breast cancer patients has explained the association of this cancer type with CDK12, PAK1, PTK2, RIPK2, and TLK2 amplicons, and highlighted the overexpression of EGFR following the loss of CETN3 and SKP1 [22]. There have been tremendous advances in the progress of tumor metabolites. The consumption and release (CORE) profiles of 219 metabolites from NCI-60 cell lines were detected and after the integrated analysis of CORE profiles with gene expression data, it was seen that the glycine consumption and upregulation of the mitochondrial glycine biosynthetic pathway were highly correlated with the proliferation of cancer cells [23].

Drug repurposing (also known as ‘new uses for old drugs’) is a strategy to identify new uses for approved or investigational drugs that are outside the scope of the original medical indication. Drug repurposing is also known as drug repositioning, drug reprofiling, indication expansion, or indication shift. It involves establishing new medical uses for already known drugs, including approved, discontinued, shelved, and experimental drugs. Although this strategy is far from new, it has gained considerable attention and momentum in the last decade. About one-third of the approvals in recent years correspond to drug repurposing; repurposed drugs currently generate around 25% of the pharmaceutical industry’s annual revenue. Many drug repurposing initiatives have been undertaken by public and nonprofit organizations. New therapeutic uses for existing compounds were initiated by the NIH–National Center for Advancing Translational Sciences in partnership with several pharmaceutical companies. The main advantages of these drugs were that they have already proven to be sufficiently safe in preclinical models or in early human trials so they are less likely to fail, at least from the safety point for the efficacy of drugs. Also, as approved drugs have successfully passed clinical trials and regulatory security, and have already undergone post-marketing surveillance, it is easy to reuse them for other diseases.

Many repurposed drugs, such as sildenafil, minoxidil, aspirin, and valproic acid are used, relying on the already known drug (e.g. an off-target adverse effect) pharmacology to solve a clinical problem for another purpose. In recent years, the drug discovery community has implemented an organized, systematic, data-driven drug repurposing approach which has integrated computational resources. Among them, are the signature matching of transcriptomic or proteomic data; similarity search approximations; structure/ligand-based virtual screenings; and systematic analysis of electronic health records, clinical trial, and postmarketing surveillance data. In the field of precision oncology, multiomic approaches have improved understanding of tumor biology and increased treatment opportunities. The ACNS02B3 brain tumor biology study, led by the Children’s Oncology group across several institutions has successfully expanded molecular profiling beyond genomics. Beyond single gene analyses, mutational signatures, RNA-based gene expression profiling, immunophenotyping, and TMB determination have proven to be useful prognostic and predictive biomarkers for response to anticancer therapies, but it is still not clear whether it will produce more successful treatment opportunities. The application of molecular profiling in clinics still faces several challenges. Data interpretation of patients with complex genomics is a big challenge. The molecular tumor board (MTB) has been

constructed to fully exploit the potential of NGS-driven therapy. MTB brings the interdisciplinary expertise for advance stage cancers from all over world so that the advice can be used for the treatment of cancer patients. The multi-disciplinary teams include oncologists, research scientists, bioinformaticians, pathologists, medical geneticists, genetic counselors, and genomicists, among others. The patient's clinical, pathologic, and molecular information is examined and, based on previous treatments and reviews of available resources for similar cases, a consensus is reached over possible treatment suggestions [24]. The interdisciplinary teams of MTB result in significant changes in treatment decisions [25]. The impact of MTB on outcomes has not yet been studied in depth, but they can help to identify patients for clinical trials, educate patients about their cancers, facilitate collaboration, and ensure that the tests and treatment of patients can be carried out in a uniform and consistent manner, based on clinical guidelines and the best available evidence. The application of AI to precision oncology is still in its infancy. Many proof-of-concept studies have been witnessed that offer a glimpse of what the NGS of precision oncology could look like Though there remain many challenges to be overcome for AI to make a mark in medicine, expectations are justifiably high, but true progress can only come from a deeper understanding of the discussed limitations and how to rectify them in an efficient manner. We look forward to seeing how AI may enhance precision oncology approaches and improve patient care around the world in the future [26]

In the coming chapters, we will cover the benefits and challenges in NGS technologies which will, over time, achieve long-term benefits for cancer patients including early prognosis, diagnosis, significant treatment, and overall survival.

References

[1] Li Y, Zheng Q, Bao C, Li S, Guo W, Zhao J, Chen D, Gu J, He X, Huang S. Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. Cell Res. (2015); 25(8):981-4. DOI: 10.1038/cr.2015.82.

[2] Zhang, L., Chung, B.Y., Lear, B.C., Kilman, V.L., Liu, Y., Mahesh, G., Meissner, R.A., Hardin, P.E., Allada, R. DN1(p) circadian neurons coordinate acute light and PDF inputs to produce robust daily behavior in Drosophila. Curr. Biol. (2010); 20(7): 591-599. DOI: 10.1016/j.cub.2010.02.056.

[3] Van Emon JM. The Omics Revolution in Agricultural Research. J Agric Food Chem. (2016); 64(1):36-44. DOI: 10.1021/acs.jafc.5b04515.
[4] Rai KM, Thu SW, Balasubramanian VK, Cobos CJ, Disasa T and Mendu V Identification, Characterization, and Expression Analysis of Cell Wall Related Genes in Sorghum bicolor (L.) Moench, a Food, Fodder, and Biofuel Crop. Front. Plant Sci. (2016); 7,1287. DOI: 10.3389/fpls.2016.01287.
[5] Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. Genome Biol (2017); 18, 83. DOI: https://doi.org/10.1186/s13059-017-1215-1.
[6] Avasthi A, Basu D, Subodh BN, Gupta PK, Sidhu BS, Gargi PD, Sharma A, Ghosh A, Rani P. Epidemiology of substance use and dependence in the state of Punjab, India: Results of a household survey on a statewide representative sample. Asian J Psychiatr. (2018); 33:18-29. DOI: 10.1016/j.ajp.2018.02.017.
[7] Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. Cell (2012) 149, 979–993. DOI: 10.1016/j.cell.2012.04.024
[8] Robinson TN, Banda JA, Hale L, Lu AS, Fleming-Milici F, Calvert SL, Wartella E. Screen Media Exposure and Obesity in Children and Adolescents. Pediatrics. (2017); 140(Suppl 2):S97-S101. DOI: 10.1542/peds.2016-1758K.
[9] Maes, J. et al. An indicator framework for assessing ecosystem services in support of the EU Biodiversity Strategy to 2020, Ecosystem Services. (2016); 17, 23. DOI:10.1016/j.ecosar.2015.10.023.
[10] Abbatiello S, Ackermann BL, Borchers C, Bradshaw RA, Carr SA, Chalkley R, Choi M, Deutsch E, Domon B, Hoofnagle AN, Keshishian H, Kuhn E, Liebler DC, MacCoss M, MacLean B, Mani DR, Neubert H, Smith D, Vitek O, Zimmerman L. New Guidelines for Publication of Manuscripts Describing Development and Application of Targeted Mass Spectrometry Measurements of Peptides and Proteins. Mol Cell Proteomics. (2017); 16(3):327-328. DOI: 10.1074/mcp.E117.067801.
[11] Bertini I, Calabrò A, De Carli V, Luchinat C, Nepi S, Porfirio B, Renzi D, Saccenti E, Tenori L. The metabonomic signature of celiac disease. J Proteome Res. (2009); 8(1):170-7. DOI: 10.1021/pr800548z.
[12] Lin FR, Albert M. Hearing loss and dementia – who is listening? Aging Ment Health. (2014); 18(6):671-673. DOI:10.1080/13607863.2014.915924
[13] Veselkov, K. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. (2019); 9:9237. DOI: https://doi.org/10.1038/s41598-019-45349-y.

[14] Quezada, H. et al. Omics-based biomarkers: current status and potential use in the clinic. Boletín Médico del Hospital Infantil de México, (2017) 74(3), 219-226. DOI:
https://doi.org/10.1016/j.bmhime.2017.11.030.
[15] Adashek JJ, Janku F, Kurzrock R. Signed in Blood: Circulating Tumor DNA in Cancer Diagnosis, Treatment and Screening. Cancers (Basel). (2021);13(14):3600. DOI:10.3390/cancers13143600.
[16] Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol. (2013) 9(4):232-240.
DOI:10.1038/nchembio.1199.
[17] Davis RL. Mechanism of Action and Target Identification: A Matter of Timing in Drug Discovery [published online ahead of print, 2020. iScience. (2020) 23(9):101487. DOI:10.1016/j.isci.2020.101487.
[18] Heo YJ, Hwa C, Lee GH, Park JM, An JY. Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. Mol Cells. (2021) 44(7):433-443.
DOI:10.14348/molcells.2021.0042.
[19] Chang HS, Lin CH, Chen YC, Yu WC. Using siRNA technique to generate transgenic animals with spatiotemporal and conditional gene knockdown. Am J Pathol. (2004) 165(5):1535-1541. DOI:10.1016/S0002-9440(10)63411-6.
[20] Leung RK, Whittaker PA. RNA interference: from gene silencing to gene-specific therapeutics. Pharmacol Ther. (2005)107(2):222-239.
DOI:10.1016/j.pharmthera.2005.03.004.
[21] Takebe N, Miele L, Harris PJ, Jeong W, Bando H, Kahn M, Yang SX, Ivy SP. Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. Nat Rev Clin Oncol. (2015);12(8):445-64. DOI: 10.1038/nrclinonc.2015.61.
[22] Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K. L., … NCI CPTAC (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature, 534(7605), 55–62.
DOI:https://doi.org/10.1038/nature18003.
[23] Jain M, Nilsson R, Sharma S, Madhusudhan N, Kitami T, Souza AL, Kafri R, Kirschner MW, Clish CB, Mootha VK. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science. (2012);336(6084):1040-4. DOI: 10.1126/science.1218595.

[24] Ungprasert P, Ryu JH, Matteson EL. Clinical Manifestations, Diagnosis, and Treatment of Sarcoidosis. Mayo Clin Proc Innov Qual Outcomes. (2019) 3(3):358-375. DOI:10.1016/j.mayocpiqo.2019.04.006.
[25] Kim JY, Kronbichler A, Eisenhut M, et al. Tumor Mutational Burden and Efficacy of Immune Checkpoint Inhibitors: A Systematic Review and Meta-Analysis. Cancers (Basel). (2019) 11(11):1798. DOI:10.3390/cancers11111798.
[26] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. (2019) 6(2):94-98. DOI:10.7861/futurehosp.6-2-94.

CHAPTER 2

TRANSCRIPTOMICS AND RNA SEQUENCING

Contents

2.1 Introduction

The revolution in genomic research comes with the development of next-generation sequencing (NGS), massively parallel or deep sequencing that describes a DNA sequencing technology. Next-generation sequencing (NGS) technologies are seen as the next step in the evolution of DNA sequencing through the generation of thousands or even millions of DNA sequences in a short time. The relatively fast emergence and success of NGS in research has developed the field of genomics and medical diagnosis. With NGS, the old traditional diagnostic model has changed to one precision medicine model, leading to more accurate diagnosis of human diseases. Using NGS, an entire human genome can be sequenced in a single day speeding up the selection of molecular target drugs for individual treatment. The application of NGS includes variant detection,

whole-exome sequencing (WES), whole-genome sequencing (WGS), custom panels (multi-gene), RNA-seq, and epigenetics among others. In this chapter we have included a detailed description of NGS (till fourth generation sequencing) that could help, scientists, researchers, and healthcare professionals to understand how to translate the genomic data into genomic medicine.

2.2 History

The concept 'inborn error of metabolism' introduced by Garrod in 1908 has changed the areas of biochemistry, genetics, and medicine [1]. His principal contribution includes the understanding of the relationship between gene-enzyme mechanisms which is the molecular basis of genetic diseases. Though his research has been superseded by the latest advances such as RNA splicing, RNAi, and others, his contributions allowed researchers to understand how changes in DNA sequence could cause genetic disease. All these findings became the basic concepts in the understanding of human DNA sequence and mutations. In 1960s, the search began to ascertain the nucleotide sequence of DNA with several studies that demonstrated new methods with different strategies [2–6].

2.2.1 First generation sequencing

In 1977, Sanger developed the 'chain-termination' method that launched a new era for first generation sequencing to sequence DNA. In this method, dideoxynucleotides (ddNTPs) were used. These are deoxynucleotide analogs (dNTPs) that disrupt DNA synthesis and separate different DNA fragments in a gel. These special nucleotides were radiolabeled enabling the sequence to be inferred after the disclosure of gel autoradiography [7]. Numerous modifications, such as the substitution of nucleotide radiolabeled to fluorescence, allowed the sequencing reaction to occur in one tube [8], the development of the polymerase chain reaction (PCR) [9], the separation of DNA fragments by capillary electrophoresis [10], and later the development of equipment that allowed the sequencing of more complex genomes to make the method more efficient, robust, and sensitive. Sanger is still considered the gold-standard method in diagnostics. As recent methods are not fully efficient, a lot of modifications have been done to make these techniques more efficient. Nowadays Sanger sequencing has been partly replaced by 'next-generation' sequencing (NGS) methods [11, 12]. The emergence of NGS has changed basic and applied sciences as well as clinical research. NGS allows identification of

biomarkers for early diagnosis, personalized treatments and produces millions of data with a minimum investment [11, 13].

2.2.2 Second generation sequencing

The Human Genome Project has produced 3 billion sequenced bases at the estimated cost of around $2.7 billion [14] in 13 years. The second generation of DNA sequencing was the era of parallel massive sequencing on a micro scale, which was developed by Nyrén and colleagues in 1996 with the name 'pyrosequencing method'. This technique differed substantially from previous ones because it did not use radio or fluorescence-labelled nucleotides and there was no need for an electrophoretic run. The method is based on the action of two enzymes: ATP sulfurylase and luciferase. ATP sulfurylase converts pyrophosphate released in nucleotide incorporation into an ATP molecule that is used by luciferase substrate. This process releases light signals in proportion to the amount of nucleotides incorporated, and the sequence can be determined according to the serial addition of nucleotides [15]. Improvising this technology, 'second-generation' equipment, 454 (Roche), was first developed. The DNA was binded in beads through an adapter and amplified in water-in-oil microreactors (emulsion PCR). These changes, along with the use of compartmentalized microplates and high-definition detection systems have dramatically increased the amount of sequenced DNA [16]. Other related technologies include 'Illumina' which consists of binding the DNA in a flow-cell through adapters; the parallel massive amplification occurs in clusters for each DNA strand that was originally bound in the flow-cell. This is called bridge-amplification. The advantage of this process is that it generates paired-end sequences, which make it superior to the other process since it improves the accuracy of mapping mainly in repetitive regions or where DNA rearrangements or gene fusions occur. 'Reversible terminator chemistry' is used in this method which is a modified fluorescent dNTP that reversibly blocks DNA synthesis, so that the addition of each nucleotide can be synchronized and monitored by a charge-coupled device (CCD) sensor [17]. This technique is the most accurate and has the lowest error rate of the sequencing methodologies used so far. The disadvantage is that it generally requires higher DNA concentration. SOLiD is another methodology, based on oligonucleotide ligation sequencing known as SOLiD developed by Applied Biosystems (now Thermo Fisher Scientific). The advantage of this method is the speed of the process and the low cost of the equipment, yet the disadvantage is the detection of homopolymers. Interestingly the second generation of

sequencing has a high capacity of sequencers in the generation of data in a single run and consequently the computational bioinformatics tools to analyse them. The 'Single Molecule Real Time' (SMRT) method commercialized by Pacific Biosciences is used for second-generation sequencing. The SMRT method consists of the immobilization of a single molecule in a chamber called 'zero-mode waveguide (ZMW)' where the incorporation of the fluorescent nucleotides occurs. ZMW allows the incorporation of each nucleotide to be monitored in real time and without interference from other light signals. The reads are very long (40 kb) and allow the detection of modified bases [18, 19].

Nanopore sequencing technology is a unique and scalable technology that enables direct and real-time analysis of long DNA or RNA fragments. As nucleic acids are passed through a protein nanopore, the changes are monitored by an electrical current. The resulting signal is decoded to provide the specific DNA or RNA sequence. The detection occurs as a result of differences in the current of ions generated by each nucleotide. The reads are incredibly long (500 kb), and the process is extremely fast without the need for special nucleotides. Oxford Nanopore Technologies (ONT) has used this technology to commercialize sequencers, including a portable version (MinION) that was used to sequence a mixture of bacteriophage, *E. coli*, and *mus musculus* DNA on the international space station (ISS) [20]. In spite of high error rates, it is used in the assembly of complex regions of the genome where gene fusions, large deletions and insertions, and repetitive regions occur.

2.2.3 Third generation sequencing

Third-generation sequencing allows long-read sequencing of individual RNA molecules [21]. Single-molecule RNA sequencing enables the generation of full-length cDNA transcripts without clonal amplification or transcript assembly. Third-generation sequencing is free from the shortcomings generated by PCR amplification and read mapping. It greatly reduces the false positive rate of splice sites and captures the diversity of transcript isoforms. Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing [22], Helicos single-molecule fluorescent sequencing [23], and Oxford Nanopore Technologies (ONT) nanopore sequencing [24] all comprises the single-molecule sequencing platforms. The evolution of these first, second, and third sequencing techniques is shown in Figure 1.

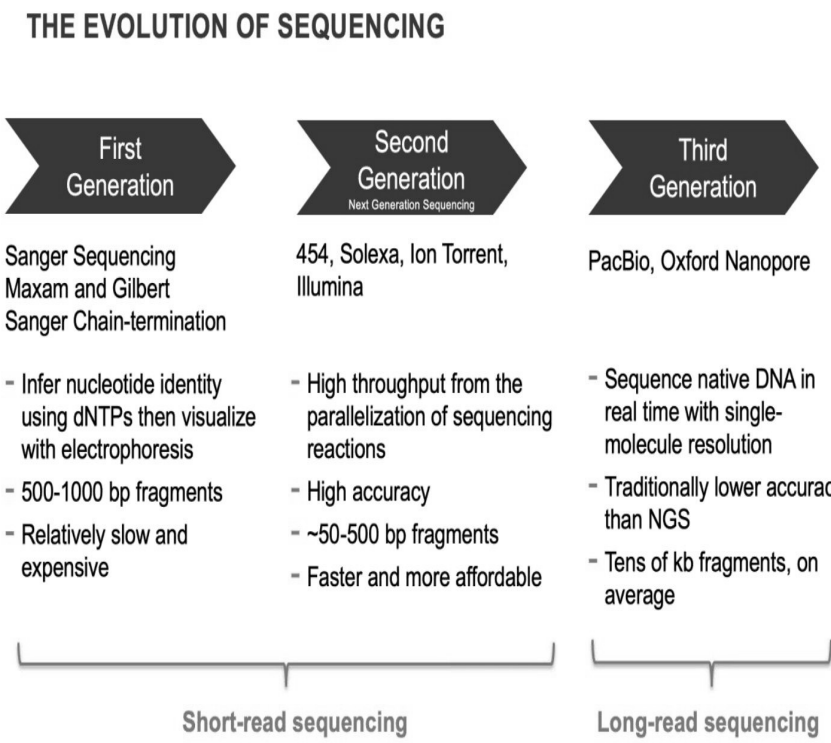

Figure 1: The evolution of first, second, and third generation sequencing technologies with their features, advantages and disadvantages. (Reused from (https://twitter.com/pacbio/status/1233091102800011266.)

2.2.4 Fourth generation sequencing

Fourth-generation RNA sequencing provides a direct *in situ* sequencing (ISS) and fluorescent ISS (FISSEQ). Although ISS and FISSEQ technologies have their own strengths in detection, they are still in the early developmental stage and many technical aspects need to be addressed before they can be applied in cancer research and clinical applications. The ISS method applied padlock probes combined with rolling circle amplification (RCA) to generate *in situ* amplified targeted sequencing libraries that are subsequently sequenced via sequencing-by-ligation NGS chemistry [25]. Through sequencing of a molecular barcode, consisting of four bases in the non-target hybridization part of the padlock

probes, the ISS method can simultaneously sequence up to 256 unique transcripts. As this method uses target specific padlock probes to create rolling circle amplification products, it is used only for sequencing known genes, such as gene panels. In contrast, the fluorescent *in situ* sequencing (FISSEQ) method uses random hexamers with a sequencing primer tag to initiate *in situ* real time (RT). Unlike cDNA in ISS, the resultant cDNAs are circularized using CircLigase. During RT, dUTP is introduced and the cDNAs are crosslinked to tissue with the reagent BS (PEG) [26] to prevent diffusion of the cDNAs. After rolling circle amplification (RCA), the products are sequenced by using the same sequencing by ligation techniques. By applying FISSEQ with a 30-base read length, 156,762 reads covering 8,102 annotated genes in human primary fibroblasts were obtained [27].

Compared to ISS methods, FISSEQ generates random libraries and allows an unbiased analysis of all cellular transcripts at a single-cell resolution. Since the majority of sequenced molecules are rRNAs, the number of transcripts detected in each cell is low [28]. ISS technology uses targeted gene panels and thus the sensitivity of ISS is around two orders of magnitude higher than that of FISSEQ for any given gene [29].

The main bottlenecks of these technologies are tissue preparation, optimized methods for improving efficiency, computational tools, and imaging scale. Fourth-generation RNAseq can potentially become a straightforward method for high-throughput spatial transcriptomic analysis in the years ahead provided that the technical obstacles can be handled in a proper manner.

2.3 Transcriptomics

The earliest sequencing-based transcriptomic methods, using serial analysis of gene expression (SAGE), worked on Sanger sequencing of concatenated random transcript fragments [30]. Basically, ‘Transcriptome’ was first used in the 1990s [31, 32]. In 1995, transcripts were quantified by matching the fragments to known genes. A variant of SAGE using high-throughput sequencing techniques, called digital gene expression analysis (DGEA), was also used [33–34]. However, these methods were superseded by high throughput sequencing of entire transcripts, as the high-throughput sequencing provides additional information on transcript structure, e.g., splice variants [34].