Article title: Translation of Genomics into Precision Medicine with Artificial Intelligence: A Narrative Review on Applications, Challenges and Future Perspectives

Authors: Arin Natania. S[1]

Affiliations: Doctor of Pharmacy, College of Pharmacy, Sri Ramakrishna Institute of Paramedical Sciences, Coimbatore, Dr. MGR Medical University, Chennai.[1]

Orcid ids: 0000-0002-3611-6078[1]

Contact e-mail: arinnatania@gmail.com

License information: This work has been published open access under Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Conditions, terms of use and publishing policy can be found at https://www.scienceopen.com/.

Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.

DOI: 10.14293/S2199-1006.1.SOR-.PP75WQS.v1

Preprint first posted online: 25 July 2020

Keywords: Genomics, Natural language processing, Precision Medicine, Artificial Intelligence, next generation sequencing, digital healthcare, cancer care
ARTIFICIAL INTELLIGENCE- A BOON!

- AI is the combination of theories, algorithms, and computing frameworks, facilitating various tasks that require human intelligence such as reasoning, decision-making, speech recognition, language understanding, and visual perception.
- It is one term that encompasses numerous methods such as logic (rule-based), machine learning (ML), deep learning, NLP, and computer vision. AI can help to significantly speed up the process of analyzing vast amounts of data, leverage patterns in the data, and result in faster and better advised decision-making. Sophisticated predictive models are built using algorithms that mine the patterns from data and predict outcomes.
- As the availability of data in a domain increases, so does the adoption and utilization of such disruptive technologies. With the advent of Big Data and the ever-increasing storage and computing power, the challenge has shifted from collecting data to turning it into meaningful and actionable insights.
**PHARMACOGENOMICS AND PRECISION MEDICINE**

- **Pharmacogenomics** is the study of how genes affect a person's response to drugs. This relatively new field combines pharmacology (the science of drugs) and genomics (the study of genes and their functions) to develop effective, safe medications and doses that will be tailored to a person's genetic makeup.

- A form of medicine that uses information about a person’s own genes or proteins to prevent, diagnose, or treat disease. In cancer, precision medicine uses specific information about a person’s tumor to help make a diagnosis, plan treatment, find out how well treatment is working, or make a prognosis. Examples of precision medicine include using targeted therapies to treat specific types of cancer cells, such as HER2-positive breast cancer cells, or using tumor marker testing to help diagnose cancer. Also called personalized medicine.
HOW NEXT-GENERATION SEQUENCING IS CHANGING THE LANDSCAPE OF CANCER GENOMICS

- Next-generation sequencing (NGS) is being applied broadly as a valuable method for gaining insights into the genomic profile of a tumour.
- Cancer panels are designed specifically to detect clinically relevant somatic mutations with high confidence. Germline mutations in cancer-predisposing genes such as BRCA1/2 are also detected to assess cancer risk.
- In 2017, the FDA approved several NGS-based panels related to oncology: Oncomine Dx Target Test, Praxis Extended RAS Panel, MSK-IMPACT, and FoundationOne CDx. Recent FDA approval of NTRK gene fusions for tumoragnostic indications also expands the clinical utilization of NGS (Larotrectinib FDA approval).
- Liquid biopsy holds great promise due to its noninvasive nature.
- Cell-free DNA (cfDNA) released by dying tumor cells, cell-derived vesicles termed exosomes, and circulating tumor cells (CTCs), which shed from the tumor and enter the vasculature system, are often used as a source for tumor DNA.
- The Cancer Genome Atlas (TCGA) project highlights how NGS screens can facilitate the discovery of novel oncogenic mechanisms and patient stratification.
- In a recent study, the regulatory role of F-box/WD repeat-containing protein 7 (Fbw7) in cancer cell oxidative metabolism is discovered (Davis et al. 2018) using ML algorithms.
- Finally, NGS supports the discovery of novel biomarkers such as mutation signatures and tumor mutational burden (TMB).
- TMB has been shown to be an effective biomarker for predicting the response to immuno-therapy—an innovative area of research that can use the body’s own immune system to fight cancer.

CHALLENGES IN CANCER GENOMICS DATA INTERPRETATION

- First, combing data profiles at various levels would result in high dimensionality with large number of covariates. Data sparsity from high dimensionality combined with high heterogeneity from diverse types of data imposes a significant difficulty in integrative analyses.
- Second, better standards for data generation and reporting are needed to facilitate data integration and to reduce bias. Sample acquisition and preparation procedures need to be well regulated for data generation and sequencing platform, and computational pipelines need to be carefully calibrated and validated.
• Last, but not least, well-designed studies with causal inference are needed to filter out biomarkers that have strong correlative effects but no real causative effects in tumorigenesis.
• On the genetic level, the pathogenic variants could be significantly enriched in cases compared to controls and/or the variant is co-inherited with disease status within affected families.
• On the informatic level, the pathogenic variants could be found at the location predicted to cause functional disruption (for example, protein-binding region). And on the experimental level, the pathogenic variants could significantly alter levels, splicing, or normal biochemical function of the product of the affected genes.
• Finally, the cellular phenotype in patient-derived cells, model organisms, or engineered equivalents can be rescued by addition of wild-type gene product or specific knockdown of the variant allele.
• The advancement of ML technologies is bound to impact the interpretation of genomic sequencing data, which has traditionally relied on manual curation by experts in the field. These curation efforts rely on protein structure, functional studies and more recently, on “in silico” models that predict the functional impact of genetic alteration such as SIFT, PANTHER-PSEP, PolyPhen2, and others.
• Two key limitations of manually curating and interpreting the results from genomics data are scalability and reproducibility. These challenges continue to grow as more genomic data become available.

**PRECISION MEDICINE AND AI**

• With precision medicine and the advancement of NGS, genomic profiles of patients have been increasingly used for risk prediction, disease diagnosis, and development of targeted therapies.
• Gene expression is an important part of the patients’ genomic profiles, and interestingly, ML classification methods applied to gene expression data are not new.
• Historically, comprehensive gene expression analysis was done with microarrays and now with RNA-seq.
• Expression data are analyzed to identify the significant genes in the upregulated or downregulated pathways (Lyu and Haque 2018; Hwang et al. 2002), and are also trained to predict the cancer subtypes and prognosis when outcome data or diagnosis information is available (Bartsch et al. 2016; Pepke and Ver Steeg 2017).
**CHALLENGES TO AI ADOPTION IN HEALTHCARE**

- **Lack of ground truth to validate the benefit**

  The evaluation of AI accuracy is critical to help gauge how well the system performs in assisting experts, and to make AI less of a black box. In cancer genomics, variant classification, clinical relevance, literature validation, and summarization are traditionally done by human experts. To prove the usefulness of an AI application, it needs to be evaluated in comparison with human experts and not only with the other AI solutions. However, this is rarely done due to the lack of publicly accessible knowledge bases for ground truth data.

- **Transparency and reproducibility**

  AI is a hot field and its use has been claimed by many platforms and companies. However, detailed information on AI techniques and models is not clearly presented and there is considerable variability in methodologies from company to company.

- **Patient/physician education**

  Digitization of healthcare has provided the access to big data information and cognitive insights to both caregivers and patients, transforming healthcare and clinical workflows (Mesko et al. 2017). The point-of-care has shifted from the clinic and physician to the patient. The old paradigm of paternalistic physician–patient relationship has been transformed into an equal-level partnership with shared medical decision-making. Experience-based medicine has evolved into evidence-based and patient-centered approaches. Both physicians and patients need to be prepared for this revolutionary role of AI in healthcare.
Major functionalities and transparency for key players in text mining and personalized medicine field

| Players             | Functionality                                                                 | Transparency                     |
|---------------------|-------------------------------------------------------------------------------|----------------------------------|
| Blueprint genetics  | Offers single gene test, targeted variant testing or whole exome sequencing service along with interpretation | No explicit AI description       |
| Cambridge cancer genomics | Uses blood tests to guide cancer therapy                                      | No explicit AI description       |
| Deep gene           | Provides cancer-type classifier based on deep learning and somatic point mutations | Publication is available (Yuan et al. 2016) |
| Deep genomics       | Develops genetic medicines using artificial intelligence technology, with a focus on the preclinical development of oligonucleotide therapies | No detailed explanation but related publication is available (Wainberg et al. 2018) |
| DeepVariant         | Analysis pipeline using a deep neural network to call genetic variants from NGS DNA data | Available in GitHub https://github.com/google/deepvariant | No explicit AI description |
| Genomeron           | Genomic search engine and database to provide disease-gene-variant relationships from the full text of the scientific literature for gene and variant interpretation |                                   |
| Genoox              | Fully customized platform for genetic applications including primary, secondary and tertiary analyses | No explicit AI description but related publication is available (Stajkovska et al. 2016) |
| Litterome           | Automatic curation systems to extract genomic knowledge from PubMed articles to facilitate browsing, searching, and reasoning | Publications are available (Poon et al. 2014, 2015) |
| Perthera            | Manage process from tumor testing through Perthera Report to provide cancer patients and physicians with therapeutic options ranked by the probability of outcome | No explicit AI description       |
| Sophia Genomics     | Provides NGS data analysis to detect, annotate and pre-classify genomic variants associated to multiple disorder areas | No explicit AI description       |
| Watson for Genomics | Provides in-depth clinical interpretation of the genetic alterations in the sample automatically, enabling clinical decision-making for personalized cancer care | No explicit AI description but related publication is available (Patel et al. 2018) |
| WuXi NextCODE       | Uses genomics to identify the underlying biology and advance the scientific understanding of disease and propel the next generation of transformative therapies | No explicit AI description but related publication is available (Zhang et al. 2018) |

For each company, the main functionality and transparency are summarized.

**PERSPECTIVES: 5 YEARS DOWNHILL**

- We have entered the advent of an era in which AI can help across the medical continuum from research to prognosis, therapy, and post cancer treatment care. AI will remain the main driver to healthcare transformation towards precision medicine. While digital health has become essential for providing best practice in healthcare, it raises some unprecedented challenges.

- Data protection, data sharing, and international standardization will be addressed and regulated (Center for Data Innovation). All those unprecedented challenges digital health poses should be addressed to ensure AI safety and beneficial impact to healthcare.
REFERENCES

- Xu 1 et al., Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Human Genetics (2019) 138:109–124. https://doi.org/10.1007/s00439-019-01970-5
- Abadi S, Yan WX, Amar D, Mayrose I (2017) A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput Biol13:e1005807. https://doi.org/10.1371/journal.pcbi.1005807
- Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, Bach PB, Murphy SB (2010) Rapid-learning system for cancer care. J Clin Oncol 28:4268–4274. https://doi.org/10.1200/JCO.2010.28.5478
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248
- Alicante A, Corazza A, Isgro F, Silvestri S (2016) Unsupervised entity and relation extraction from clinical records in Italian. ComputBiol Med 72:263–275. https://doi.org/10.1016/j.compbiomed.2016.01.014
- Alves R, Rodriguez-Baena DS, Aguilar-Ruiz JS (2009) Gene association analysis: a survey of frequent pattern mining from gene expression data. Brief Bioinform 11:210–224. https://doi.org/10.1093/bib/bbp042
- Amendola LM et al (2016) Performance of ACMG-AMP variant interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. Am J Hum Genet 99:247. https://doi.org/10.1016/j.ajhg.2016.06.001
- Antaki D, Brandler WM, Sebat J (2018) SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. Bioinformatics 34:1774–1777. https://doi.org/10.1093/bioinformatics/btx813
- Aravanis AM, Lee M, Klausner RD (2017) Next-generation sequencing of circulating tumor DNA for early cancer detection. Cell 168(4):571–574. https://doi.org/10.1016/j.cell.2017.01.030
- Bao L, Cui Y (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21:2185–2190. https://doi.org/10.1093/bioinformatics/bti365
- Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA (2011) PESCADOR, a webbased tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinform 12:435. https://doi.org/10.1186/1471-2105-12-435
- Bartsch G Jr et al (2016) Use of artificial intelligence and machine learning algorithms with gene expression profiling to predict recurrent nonmuscle invasive urothelial carcinoma of the bladder. J Urol 195(2):493–498. https://doi.org/10.1016/j.juro.2015.09.090