Peertechz

Cancer Genomics and Epigenetics
OPEN ACCESS JOURNAL
DOI: https://dx.doi.org/10.17352/cge

Mini Review

Identification and Validation of Cancer Mutations Using Computational Approaches: A Review

Janani Vijayaraj*

SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India

Received: 12 February, 2025
Accepted: 27 February, 2025
Published: 28 February, 2025

***Corresponding author:** Janani Vijayaraj, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, India, E-mail: vijayarajbv@rediffmail.com

Keywords: Driver mutations; Genomic data; Variant calling; Machine learning; Network-based approaches; Multi-omics data

https://www.cancerresgroup.com

Check for updates

Abstract

Cancer is a genetic disease driven by somatic mutations, with a subset of these mutations acting as drivers to promote tumorigenesis. Identifying and validating these driver mutations is essential for understanding cancer biology and developing targeted therapies. With the explosion of genomic data from large-scale sequencing projects, computational approaches have become indispensable tools for analyzing these data, predicting functional mutations, and distinguishing driver mutations from passengers. This review provides an overview of key computational methods for cancer mutation analysis, including variant calling, driver mutation identification, machine learning, and network-based approaches. It discusses current challenges, the application of these methods, and future directions, emphasizing the integration of multi-omics data and Artificial Intelligence (AI) to drive advancements in cancer research and personalized medicine.

Introduction

Cancer arises from the accumulation of somatic mutations in the genome, which can disrupt critical cellular pathways and drive tumorigenesis. While the majority of these mutations are 'passenger mutations' that do not contribute to cancer progression, a smaller subset are 'driver mutations' that confer a selective growth advantage to cancer cells. Identifying and validating these driver mutations is pivotal for unraveling the molecular mechanisms underlying cancer and for developing precision therapies [1].

The TP53 and ATM genes are essential for the development of cancer, and their mutations have substantial implications for DNA repair mechanisms and tumor suppression. The TP53 gene, often referred to as the "guardian of the genome," encodes a tumor suppressor protein responsible for DNA repair, apoptosis, and cell cycle regulation. Similarly, ATM encodes a serine/threonine kinase critical for the DNA Damage Response (DDR) pathway. Mutations in these genes are frequently observed in various malignancies, such as lung cancer, and serve as biomarkers for diagnosis, prognosis, and therapy [2]. This mini-review provides a comprehensive examination of these genes, with a focus on variant calling, driver mutation identification, network-based approaches, structural bioinformatics, pan-cancer analysis, machine learning predictions, and non-coding mutations. We explore how these elements contribute to our understanding of cancer genomics and their potential for therapeutic targeting.

The advent of high-throughput sequencing technologies and large-scale genomic initiatives, such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), has generated vast amounts of cancer genomic data [3]. Computational approaches have emerged as powerful tools for analyzing these datasets, enabling the identification of driver mutations, assessing their functional impact, and exploring their role in cancer biology [4]. This review presents a comprehensive overview of the computational methods used in cancer mutation analysis, their current applications, and future directions.

Methods

Variant calling and annotation

The first step in cancer mutation analysis involves

Citation: Vijayaraj J. Identification and Validation of Cancer Mutations Using Computational Approaches: A Review. Cancer Genom Epigenet. 2025;1(1):001-004. Available from: https://dx.doi.org/10.17352/cge.000001

identifying somatic mutations from Next-Generation Sequencing (NGS) data. Widely used tools such as GATK (Genome Analysis Toolkit) and VarScan facilitate variant calling by detecting mutations in raw sequencing data. Once identified, mutations are annotated using tools like ANNOVAR and Ensembl VEP, which provide functional predictions, clinical relevance, and insights into mutation consequences on gene expression and protein function [5].

Accurate variant calling requires careful consideration of the cancer type and study design. For example, GATK is widely used for its robust handling of high-depth sequencing data, while MuTect2 is preferred for its sensitivity in detecting low-frequency mutations in heterogeneous tumor samples [4]. Resources like the Cancer Genome Atlas (TCGA) and the Catalogue of Somatic Mutations in Cancer (COSMIC) provide guidelines and benchmarks for tool selection based on cancer type and sequencing platform [3,6]. Additionally, the choice of annotation tools should consider the specific biological questions being addressed. For instance, ANNOVAR is well-suited for clinical annotation, while Ensembl VEP excels in functional prediction and pathway analysis [4,7].

Driver mutation identification

A critical challenge in cancer genomics is distinguishing driver mutations from passenger mutations. Tools such as MutSigCV, OncodriveCLUST, and IntOGen identify significantly mutated genes and driver mutations by analyzing mutation clustering patterns, recurrence across samples, and their functional impact on tumorigenesis [4].

These tools differ in their underlying algorithms and applications. MutSigCV uses a statistical model to identify genes with more mutations than expected by chance, making it suitable for large-scale pan-cancer studies [4]. OncodriveCLUST focuses on identifying mutations that cluster in specific protein domains, which is particularly useful for studying oncogenes [8]. IntOGen integrates multiple data types, including mutation frequency and functional impact scores, to prioritize driver mutations [9]. The choice of tool depends on the study's focus: MutSigCV for broad discovery, OncodriveCLUST for domain-specific insights, and IntOGen for integrative analysis.

Machine learning and deep learning

Machine Learning (ML) methods have become increasingly effective in predicting the functional impact of mutations. Classical models like random forests and Support Vector Machines (SVMs) leverage features such as protein structure, evolutionary conservation, and gene expression profiles. More recently, deep learning models have shown remarkable promise [3].

For instance, random forests are effective for handling high-dimensional data and identifying non-linear relationships, while SVMs excel in classifying mutations based on their functional impact [4]. However, deep learning models, such as those developed by Poplin, et al. [5], outperform traditional ML methods by capturing complex patterns in large datasets [7]. These models integrate genomic, transcriptomic, and proteomic data to improve mutation prediction accuracy, making them particularly valuable for personalized cancer therapy [4].

Network-based approaches

Network-based tools such as HotNet2 and DawnRank analyze mutations within the context of biological networks (e.g., protein-protein interaction networks) to identify mutation hotspots and dysregulated cellular pathways [4].

The integration of network data enhances the identification of critical signaling pathways by revealing interactions between mutated genes and their functional partners. For example, HotNet2 identifies subnetworks with significant mutation enrichment, while DawnRank ranks genes based on their influence within the network [10]. These approaches provide a more holistic understanding of cancer biology compared to gene-centric methods, uncovering potential therapeutic targets that might otherwise be overlooked [11].

Structural bioinformatics

Understanding the impact of mutations on protein structure and function is crucial for cancer research. Structural bioinformatics tools like FoldX, Rosetta, and I-Mutant predict how mutations may disrupt protein stability, folding, or interactions [12].

The functional impact of particular alterations is clarified through structural analysis of TP53 and ATM mutations. The visualization and modeling of protein structures are facilitated by tools such as PyMOL and SWISS-MODEL [13]. ATM mutations can impair kinase activity, resulting in deficient DDR signaling, while certain missense mutations in TP53 destabilize its DNA-binding domain [4]. Recent studies have demonstrated these findings [5].

Results and discussion

Pan-cancer analysis of driver mutations

Large-scale cancer projects like TCGA and ICGC have enabled the identification of common and unique driver mutations across different cancer types [4].

Recent initiatives such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) have expanded our understanding of tumor heterogeneity. PCAWG, for example, has identified non-coding driver mutations and structural variants across 2,658 cancer genomes, while CPTAC integrates proteomic data to link mutations to functional protein changes [6]. These resources provide valuable datasets for cross-cancer comparisons and the identification of shared therapeutic targets [4].

Machine learning for mutation prediction

Machine learning models have been widely applied to predict the functional consequences of cancer mutations [4].

Citation: Vijayaraj J. Identification and Validation of Cancer Mutations Using Computational Approaches: A Review. Cancer Genom Epigenet. 2025;1(1):001-004. Available from: https://dx.doi.org/10.17352/cge.000001

For instance, Tokheim, et al. [14] developed machine learning models that predict the functional impact of mutations on protein structure, while Poplin, et al. [5] demonstrated how deep learning models improve mutation detection accuracy. These models outperform traditional methods by integrating multi-omics data and capturing complex biological relationships, offering new insights into cancer biology and personalized therapy [4,8].

Non-coding mutations

Recent attention has shifted towards non-coding mutations, which include mutations in promoters, enhancers, and other regulatory regions [9].

Non-coding regions, which were previously regarded as genomic "dark matter," are now more widely acknowledged for their involvement in cancer. Gene expression can be disrupted by mutations in regulatory elements, including enhancers and promoters [3]. FunSeq2 and FATHMM-MKL are tools that are intended to identify potentially pathogenic non-coding mutations [4]. However, the functional significance of these mutations remains difficult to interpret due to the limited number of annotations [7].

Conclusion

The identification and validation of cancer mutations using computational approaches have revolutionized our understanding of cancer biology. Leveraging large-scale genomic datasets, advanced computational methods, and interdisciplinary collaboration, researchers are uncovering novel driver mutations that could lead to the development of targeted therapies. The integration of multi-omics data and AI-driven tools offers exciting opportunities for advancing cancer research and improving patient outcomes. Moving forward, computational approaches will continue to play a pivotal role in personalizing cancer treatments, ultimately leading to more effective and tailored therapeutic strategies.

Cancer research continues to be fundamentally influenced by mutations in TP53 and ATM, which provide valuable insights into therapeutic targeting and tumorigenesis. It is imperative to combine computational methods with experimental validation in order to advance personalized cancer treatments. Future research should concentrate on the integration of multi-omic data and the utilization of advanced AI techniques to gain a more profound understanding of cancer genomics.

Acknowledgment

This manuscript was prepared with the assistance of DeepSeek-V3, an AI language model developed by DeepSeek. The tool was used to assist in drafting, organizing, and refining the content of this review. Specifically, the AI tool was utilized to:

1. Generate initial drafts of sections based on provided outlines and key points.
2. Suggest improvements to the structure, flow, and clarity of the text.
3. Provide references and citations to support the claims made in the manuscript.

After the AI-generated content was produced, I thoroughly reviewed, edited, and revised the manuscript to ensure accuracy, relevance, and alignment with the intended scope of the review. I also critically evaluated the references and ensured that the content adhered to the highest standards of academic integrity.

I assume full responsibility for the content of this publication, including the accuracy of the information presented, the validity of the references cited, and the overall quality of the manuscript. The use of AI tools was solely to enhance the efficiency of the writing process, and the final content reflects my intellectual contribution and oversight.

References

1. Dimitrakopoulos CM, Beerenwinkel N. Computational approaches for the identification of cancer genes and pathways. Wiley Interdiscip Rev Syst Biol Med. 2017;9(1):e1364. Available from: https://doi.org/10.1002/wsbm.1364
2. 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 Apr 5;173(2):371-385.e18. Erratum in: Cell. 2018;174(4):1034-1035. Available from: https://doi.org/10.1016/j.cell.2018.02.060
3. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Cancer Genome Atlas Research Network; Van Allen EM, Cherniack AD, Ciriello G, Sander C, Schultz N. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321-337.e10. Available from: https://doi.org/10.1016/j.cell.2018.03.035
4. Cava C, Sabetian S, Salvatore C, Castiglioni I. Pan-cancer classification of multi-omics data based on machine learning models. Netw Model Anal Health Inform Bioinform. 2024;13:6. Available from: http://dx.doi.org/10.1007/s13721-024-00441-w
5. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-987. Available from: https://doi.org/10.1038/nbt.4235
6. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941-D947. Available from: https://doi.org/10.1093/nar/gky1015
7. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. MC3 Working Group; Cancer Genome Atlas Research Network. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. 2018;6(3):271-281.e7. Available from: https://doi.org/10.1016/j.cels.2018.03.002
8. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2015;47(2):106-14. Available from: https://doi.org/10.1038/ng.3168
9. Ng PK, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, Sengupta S, et al. Systematic Functional Annotation of Somatic Mutations in Cancer. Cancer Cell. 2018;33(3):450-462.e10. Available from: https://doi.org/10.1016/j.ccell.2018.01.021
10. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. Available from: https://doi.org/10.1126/scisignal.2004088

Citation: Vijayaraj J. Identification and Validation of Cancer Mutations Using Computational Approaches: A Review. Cancer Genom Epigenet. 2025;1(1):001-004. Available from: https://dx.doi.org/10.17352/cge.000001

11. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Cancer Genome Atlas Research Network; Benz CC, Perou CM, Stuart JM. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929-944. Available from:https://doi.org/10.1016/j.cell.2014.06.049

12. Kandoth C, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333-339. Available from: https://www.nature.com/articles/nature12634

13. Chakravarty D, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. OncoKB: A precision oncology knowledge base. JCO Precision Oncology. 2017;2017:PO.17.00011. https://doi.org/10.1200/po.17.00011

14. Tokheim C, Karchin R. CHASMplus reveals the scope of somatic missense mutations driving human cancers. Cell Syst. 2019;9(1):9-23. Available from: https://doi.org/10.1016/j.cels.2019.05.005

Citation: Vijayaraj J. Identification and Validation of Cancer Mutations Using Computational Approaches: A Review. Cancer Genom Epigenet. 2025;1(1):001-004. Available from: https://dx.doi.org/10.17352/cge.000001